Financial Risk for Loan Approval

Zuzanna Herdzik, Julita Janik, Anna Kościelecka

February 18, 2026

1 Introduction

1.1 Description of the Dataset

This dataset contains 20,000 synthetic loan application records, generated to model financial risk and loan approval decisions. It includes detailed information about applicants' demographics, income, employment status, education, credit history, loan purpose, and other financial indicators. The data can be used both for regression tasks, such as predicting a borrower's risk score, and for binary classification, to determine the likelihood of loan approval. The attributes:

Attribute	**Description**	**Unit/Type**
`ApplicationDate`	Date when the loan application was submitted (e.g. 2022-03-15)	Date
`Age`	Applicant's age, ranging from 18 to 80 years (18-80 years)	Numeric
`AnnualIncome`	Total yearly income of the applicant, usually between 15,000 and 485,341 USD	Numeric
`CreditScore`	Numerical creditworthiness score, typically ranging from 343 to 712	Numeric
`EmploymentStatus`	Current job situation (e.g. Employed, Unemployed, Self-Employed)	Categorical
`EducationLevel`	Highest level of education attained (High School, Associate, Bachelor, Master, Doctorate)	Categorical
`Experience`	Total years of work experience (0–61 years)	Numeric

Attribute	Description	Unit/Type
LoanAmount	Amount requested by the applicant (3,674–184,732 USD)	Numeric
LoanDuration	Loan repayment period (12–120 months)	Numeric
MaritalStatus	Applicant's marital status (Single, Married, Divorced, Widowed)	Categorical
NumberOfDependents	Number of financially dependent persons (0–5)	Numeric
HomeOwnershipStatus	Indicates whether the applicant owns, rents, or has a mortgage	Categorical
MonthlyDebtPayments	Total monthly payments for existing debts (50–2,919 USD)	Numeric
CreditCardUtilizationRate	Percentage of credit limit currently in use (0–92%)	Numeric
NumberOfOpenCreditLines	Number of currently active credit accounts (0–13)	Numeric
NumberOfCreditInquiries	Number of recent credit report checks (0–7)	Numeric
DebtToIncomeRatio	Ratio of total debt to monthly income (0.0–0.9)	Numeric
BankruptcyHistory	Indicates whether the applicant has previously declared bankruptcy (1/0)	Binary
LoanPurpose	Purpose of the loan (e.g. Debt Consolidation, Home, Auto, Education)	Categorical
PreviousLoanDefaults	Whether the applicant defaulted on previous loans (1/0)	Binary
PaymentHistory	Past payment record (8-45)	Numeric
LengthOfCreditHistory	Duration of credit history (1–29 years)	Numeric
SavingsAccountBalance	Current balance of savings account (73–200,089 USD)	Numeric
CheckingAccountBalance	Current balance of checking account (24–52,572 USD)	Numeric
TotalAssets	Total value of all owned assets (2,098–2,619,627 USD)	Numeric
TotalLiabilities	Total amount of financial obligations (372–1,417,302 USD)	Numeric

Attribute	Description	Unit/Type
MonthlyIncome	Monthly income, derived from annual income (1,250–25,000 USD)	Numeric
UtilityBillsPaymentHistory	Record of utility bill payments (0.26-1)	Numeric
JobTenure	Duration of current employment (0–16 years)	Numeric
NetWorth	Total assets minus liabilities (1,000 to 2,603,208 USD)	Numeric
BaseInterestRate	Base lending rate before adjustments (0.13–0.41)	Numeric
InterestRate	Final applied interest rate for the loan (0.11–0.45)	Numeric
MonthlyLoanPayment	Monthly payment amount for the approved loan (97–10,900 USD)	Numeric
TotalDebtToIncomeRatio	Overall debt compared to income (0.02–4.65)	Numeric
LoanApproved	Loan approval outcome (1/0, output target)	Binary
RiskScore	Model-based score estimating probability of default (28.8–84, output target)	Numeric

It should be noted that there is no missing values in the dataset.

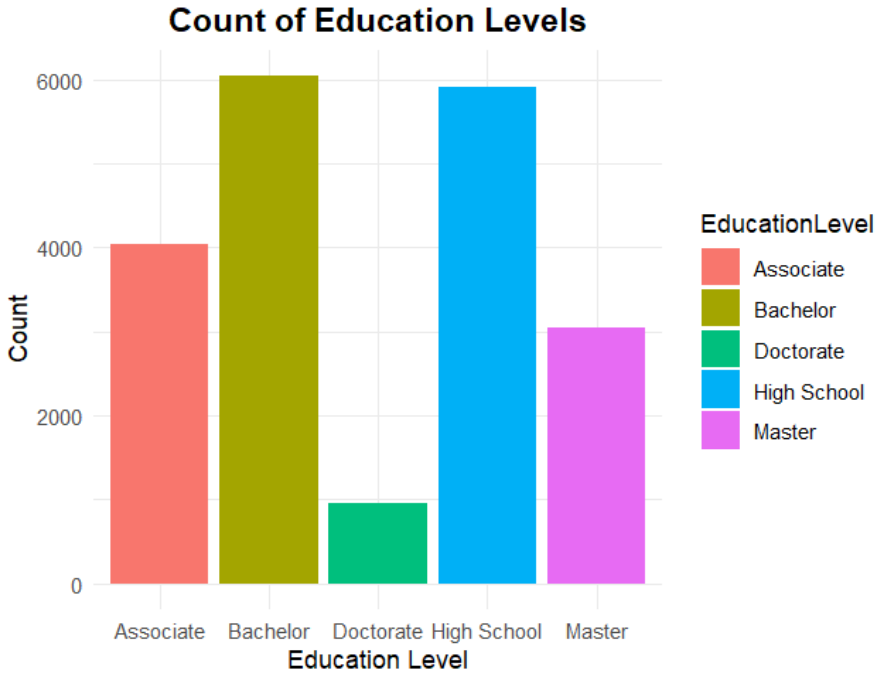
2 Data Analysis

In this part we will analyze individual variables. We will consider the impact of various variables on Loan Approval and Risk Score.

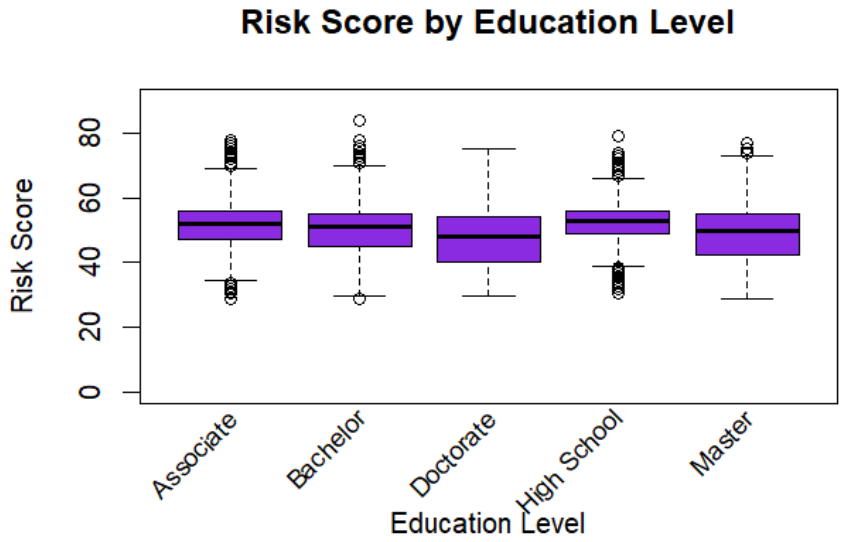
3 Education

In this chapter, we examine the potential relationship between applicant's risk score and their education level. In particular, we assume that there may be differences in risk score outcomes between applicants with higher education and those with lower levels of education.
The counts for each category are as follows:

- High School: 5908
- Associate: 4034
- Bachelor: 6054
- Master: 3050
- Doctorate: 954



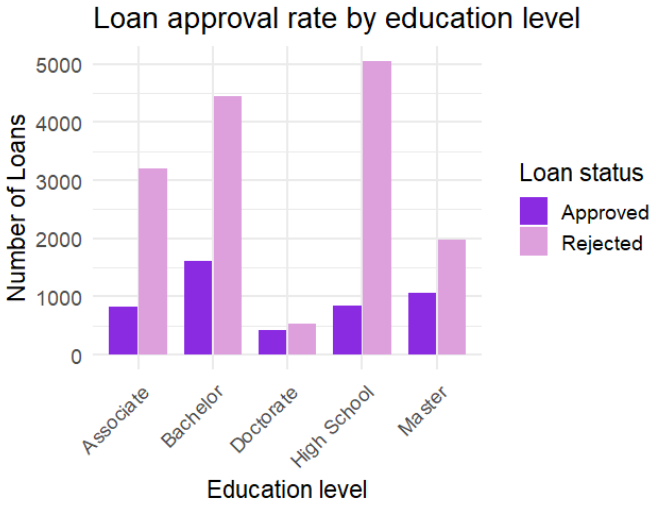
The boxplots for each group of education level:



The median allows us to draw conclusions about the performance of at least half of the applicants in each category. For instance, the highest medians are observed in the high school and associate categories.
The table below presents the mean risk scores of applicants categorized by their education level.

Education level	Associate	Bachelor	Doctorate	High School	Master
Median	52	51	48	53	50

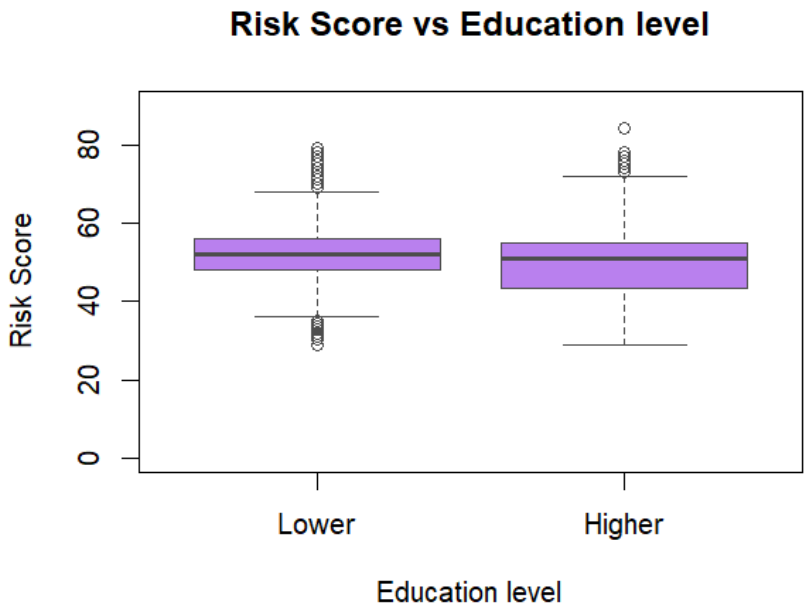
In addition, we present the loan approval status:



3.1 Impact of Education level on Risk Score

The objective of this analysis was to examine whether applicants with higher levels of education (Bachelor's, Master's, Doctorate) exhibit statistically different risk scores compared to those with lower levels of education (High School, Associate).

To present and compare the differences in risk scores between these two groups, we demonstrate the boxplot:



While the median values appear relatively close, suggesting comparable central tendencies, the overall spread of scores differs substantially. The group with higher education exhibits a wider range and greater variability.

- Mean risk score for the lower education group: $\mu_{lower} = 51.86067$
- Mean risk score for the higher education group: $\mu_{higher} = 49.6855$
- Overall mean across all applicants:: $\mu_{total} = 50.76678$

3.1.1 Hypothesis Formulation

Let μ_{lower} and μ_{higher} represent the mean risk scores for the respective groups.

- **Null hypothesis** H_0:

$$\mu_{\text{lower}} = \mu_{\text{higher}}$$

There is no statistically significant difference in average risk scores between applicants with lower and higher education levels.

- **Alternative hypothesis** H_1:

$$\mu_{\text{lower}} \neq \mu_{\text{higher}}$$

The average risk scores differ significantly depending on the level of education.

In the context of a linear regression model, this comparison can be reframed as a test of coefficient equality:

$$H_0 : \beta_1 = \beta_2 = \beta_{12} \qquad \text{vs.} \qquad H_1 : \beta_1 \neq \beta_2$$

Here, β_1 and β_2 denote the estimated effects of each education category on the predicted risk score.

Two models were considered. The first one is **Full Model**, defined by the formula:

$$\mu(R_{i1}, R_{i2}) = \beta_1 \cdot R_{i1} + \beta_2 \cdot R_{i2}$$

where:

$$R_{i1} = \begin{cases} 1 & \text{if the applicant has lower education} \\ 0 & \text{otherwise} \end{cases}$$

and

$$R_{i2} = \begin{cases} 1 & \text{if the applicant has higher education} \\ 0 & \text{otherwise} \end{cases}$$

Result:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    51.86067    0.07724  671.38   <2e-16 ***
EduGroupHigher -2.17517    0.10893  -19.97   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.702 on 19998 degrees of freedom
Multiple R-squared:  0.01955,   Adjusted R-squared:  0.0195
F-statistic: 398.8 on 1 and 19998 DF,  p-value: < 2.2e-16
```

The second one is **Reduced Model**, defined as:

$$\mu(R_{i1}, R_{i2}) = \beta_{12} \cdot (R_{i1} + R_{i2}) = \beta_{12}$$

Result:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   50.767      0.055     923   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.778 on 19999 degrees of freedom
```

3.1.2 Statistical Testing

To formally evaluate whether the level of education is associated with significant differences in applicants' risk scores, we apply two complementary statistical procedures: the **F-statistic** for model comparison and the **t-test** for mean difference assessment.

We compare two models:

- Model 0: assumes equal mean risk scores across education levels (pooled model),
- Model 1: allows separate mean estimates for lower and higher education groups.

The **F-statistic** is computed as:

$$F_{\text{obs}} = \frac{\text{SSE}_0 - \text{SSE}_1}{df_0 - df_1} \cdot \frac{df_1}{\text{SSE}_1}$$

where:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- SSE_0, SSE_1 are the sums of squared errors for Model 0 and Model 1,
- df_0, df_1 are the degrees of freedom for each model.

In our analysis:

$$F_{\text{obs}} = 398.776, \qquad F_{\text{crit}} = 3.841924 \quad (\alpha = 0.05)$$

If $F_{\text{obs}} > F_{\text{crit}}$, we reject the null model and conclude that education level significantly improves model fit.

To test whether the average risk scores differ between the two education groups, we also use the **t-test**:

$$t_{\text{obs}} = \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_0} \right)}}$$

where:

$$s^2 = w_1 \cdot \mathrm{Var}_1 + w_0 \cdot \mathrm{Var}_0$$

$$w_1 = \frac{n_1 - 1}{n_0 + n_1 - 2}, \qquad w_0 = \frac{n_0 - 1}{n_0 + n_1 - 2}$$

- $\bar{X}_1$, $\bar{X}_0$ are the sample means for higher and lower education groups,
- n_1, n_0 are the respective sample sizes,
- s^2 is the pooled variance estimate.

We compare the observed t-statistic to the critical values for a two-tailed test at significance level $\alpha = 0.05$:

$$t_{\text{obs}} = \text{-19.96938}, \qquad t_{\text{left}} = \text{-1.960083}, \qquad t_{\text{right}} = 1.960083$$

If $t_{\text{obs}} < t_{\text{left}}$ or $t_{\text{obs}} > t_{\text{right}}$, we reject the null hypothesis, indicating a statistically significant difference in mean risk scores between the two education levels.

Interpretation The results show that applicants with higher education levels usually have slightly lower risk scores than those with lower education levels. This means that people with higher education may be seen as less risky borrowers, possibly because they have more stable jobs, higher incomes, or manage their finances more carefully. Even though the difference in average scores is statistically significant, it is quite small, suggesting that education level affects the risk score only to a limited extent and that other factors also play an important role.

3.2 The analysis of Risk Score based on Interest Rate and Education Level

To examine whether the relationship between risk score and interest rate varies across education levels, so similarly to above we define two binary indicators that distinguish applicants by their educational background:

$$R_{i1} = \begin{cases} 1 & \text{if the applicant has lower education} \\ 0 & \text{otherwise} \end{cases}$$

and

$$R_{i2} = \begin{cases} 1 & \text{if the applicant has higher education} \\ 0 & \text{otherwise} \end{cases}$$

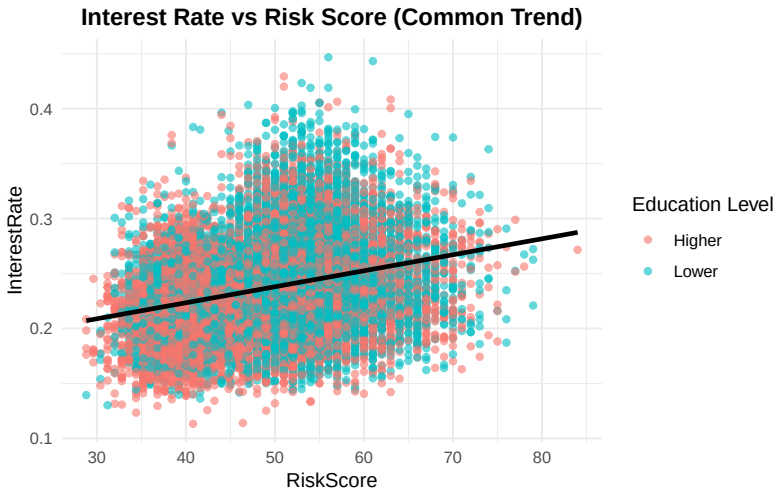
3.2.1 Modeling a Common Trend in Interest Rate by Education Level

To test whether the relationship between risk score and interest rate is consistent across education levels, we specify a simplified model that assumes a shared slope for both groups. Under the hypothesis

$$H_0 : \beta_1^1 = \beta_1^2 = \beta_1^{12}$$

we define model:

$$\mu(x_i, R_{i1}, R_{i2}) = \beta_0^1 R_{i1} + \beta_0^2 R_{i2} + \beta_1^{12} \cdot x_i$$



3.2.2 Modeling a Different Trends in Interest Rate by Education Level

Assuming that the effect of education level on risk score may follow different trends in each group, under the hypothesis

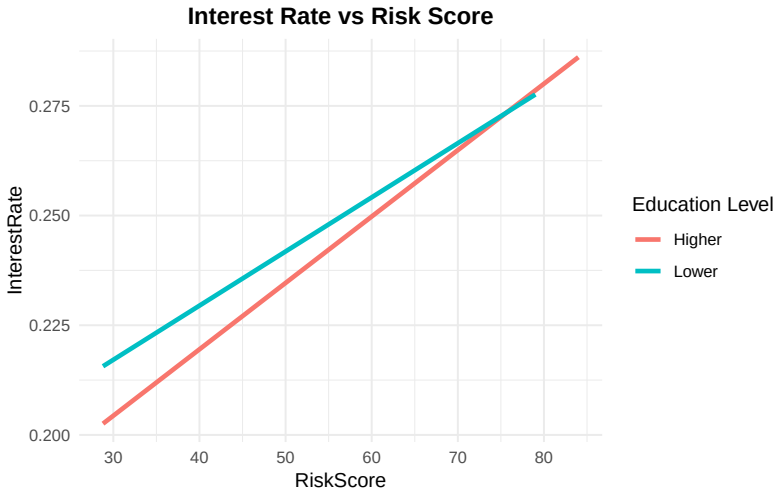
$$H_1 : \beta_1^1 \neq \beta_1^2$$

we specify the full model as:

$$\mu(x_i, R_{i1}, R_{i2}) = \beta_0^1 R_{i1} + \beta_1^1 R_{i1} \cdot x_i + \beta_0^2 R_{i2} + \beta_1^2 R_{i2} \cdot x_i$$

where

- β_0^1, β_0^2 - intercepts for the higher and lower education groups,
- β_1^1, β_1^2 - slope coefficients for the effect of x_i within each group.



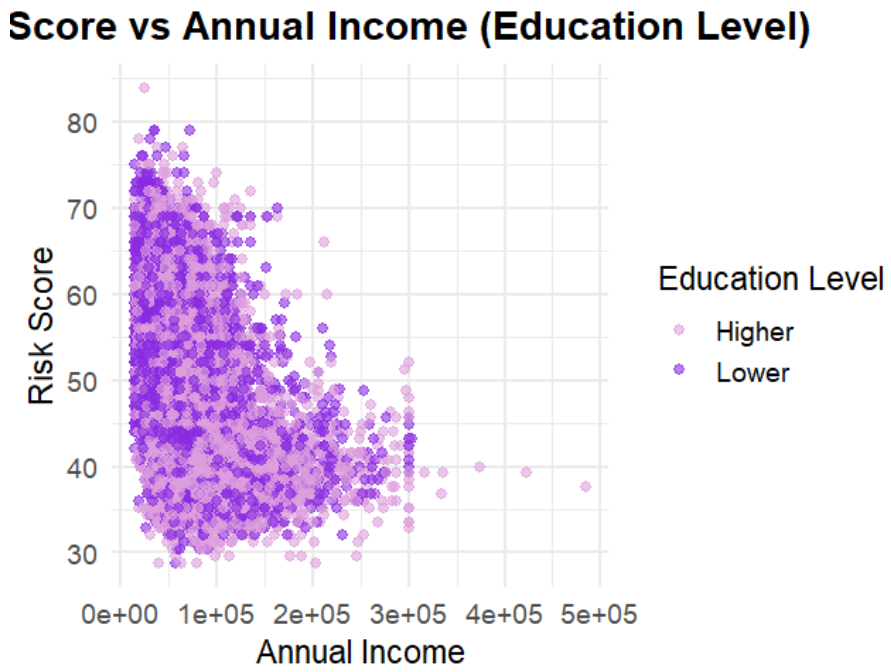
3.2.3 Conclusions

The F-statistic (13.826) exceeds the critical value ($\approx$ 3.84), indicating that the interaction between Risk Score and Education Level significantly improves model fit. Therefore, the model with distinct slopes should be preferred.

```
Analysis of Variance Table

Model 1: InterestRate ~ RiskScore + EducationCategory
Model 2: InterestRate ~ RiskScore * EducationCategory
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1  19997 32.827
2  19996 32.805  1  0.022683 13.826 0.0002011 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3.3 Analysis of Annual Income and Risk Score by Education Level

We first conduct exploratory data analysis by visualizing the relationship between applicants' annual income and their corresponding risk scores, grouped by education level (higher vs lower).



The plot suggests that there might be a weak negative relationship between annual income and risk score, as higher income is generally associated with lower risk scores.

The piecewise linear regression. Based on the plot above, we assume that there is a single break point, although the value of τ is unknown. To determine which of the potential break points is the most suitable, we will evaluate the following models for each candidate τ_j from the set of possible break points $\xi = \{15,000; \ldots; 485,341\}$:

$$\mu^{(j)}(x_i) = \beta_0^{(j)} + \beta_1^{(j)} x_i + \delta^{(j)} \Psi_{\tau_j}(x_i)$$

Next, knowing the values of SSE_j, we will choose the final value of τ as the one for which SSE_j is minimized:

$$\tau = \arg\min_{\tau_j \in \xi} SSE_j$$

The results indicate that the best value for τ is 130,587 which minimizes the SSE. Let us then consider the model:

$$E(Y_i \mid x_i) = \beta_0 + \beta_1 x_i + \delta(x_i - 130587)$$

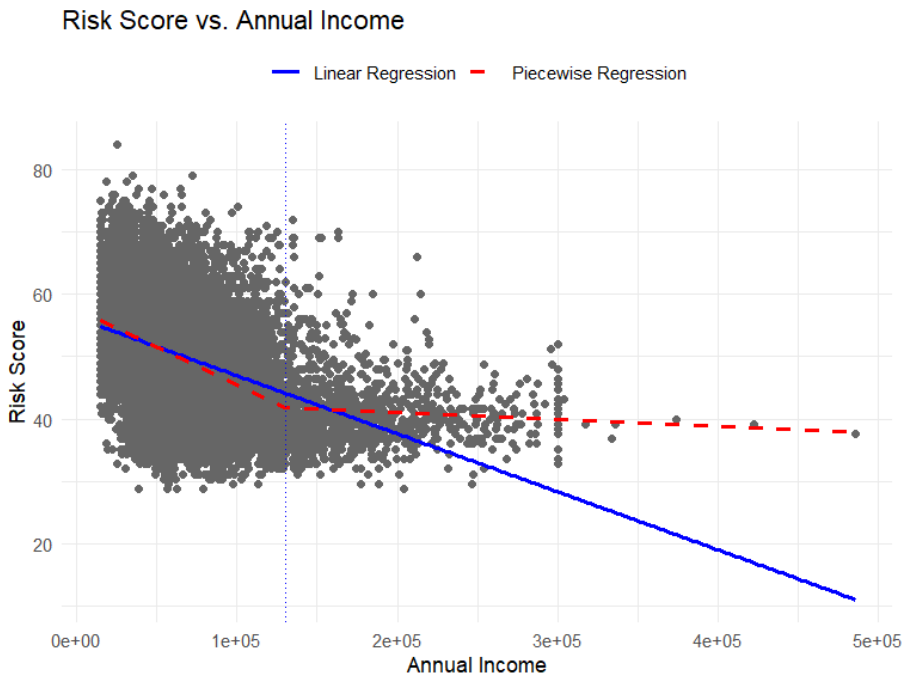
```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.764e+01  1.002e-01  575.11   <2e-16 ***
AnnualIncome -1.210e-04  1.618e-06  -74.77   <2e-16 ***
hinge         1.099e-04  4.395e-06   25.00   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.706 on 19997 degrees of freedom
Multiple R-squared:  0.2568,    Adjusted R-squared:  0.2567
F-statistic:  3455 on 2 and 19997 DF,  p-value: < 2.2e-16
```

Conclusion The slope of the line after the break point (τ = 130,587) is approximately -0.000012, which indicates that as annual income increases beyond this threshold, the risk score tends to decrease only slightly. Specifically, for individuals with income below τ, the slope is -0.000121, suggesting a stronger negative relationship between income and risk score meaning that higher income is associated with lower risk. However, after the break point, the relationship becomes much weaker, implying that increases in income above 130,587 do not significantly reduce the risk score any further.

Linear regression model Now, the model is expressed as:

$$E(Y_i \mid x_i) = \beta_0 + \beta_1 x_i$$

Let us present these regression models on the graph.



We compared the two models by examining whether the slope of the relationship after τ differs from the slope before τ.

To investigate this, we test the following hypotheses:

$$H_0 : \delta = 0$$
$$H_1 : \delta \neq 0$$

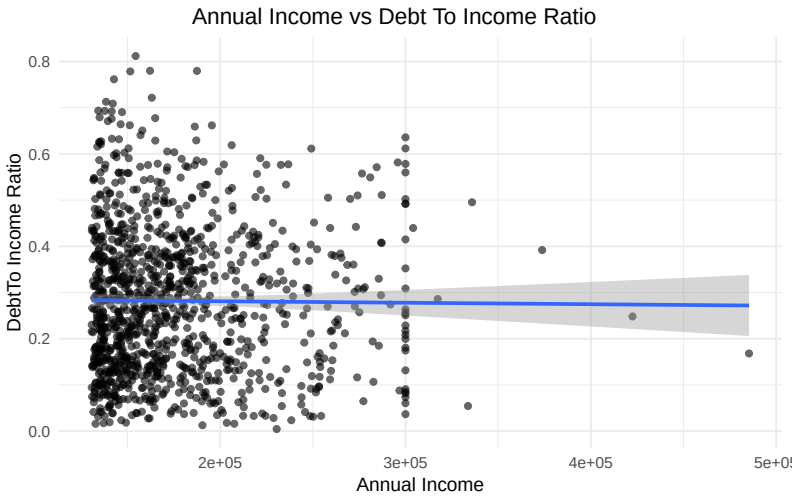
Traditionally, to compare these two models, we will use the F-statistic, determining the acceptance region for H_0 and the rejection region for H_1. The F-statistic is calculated as:

$$F_{\text{stat}} = \frac{(\text{SSE}_0 - \text{SSE}_1)/(df_0 - df_1)}{\text{SSE}_1/df_1} = 312.5068$$

which is bigger than the **threshold** value of 2.99618. This suggests that the two models differ significantly, and therefore, we choose the more complex model (the one with the spline / breakpoint).

Summary The spline model shows that income does not affect risk in the same way across the whole range. For low and medium incomes, higher income clearly goes together with lower risk scores. Once income is above roughly 130,000, the curve becomes almost flat earning more money after this point hardly changes the risk score at all.

On the plot above, we observe that while the risk score generally decreases with increasing annual income, the decline becomes noticeably less steep beyond a certain income threshold. This suggests that for wealthier individuals, higher income does not translate into proportionally lower risk. The plot below supports this interpretation: it focuses exclusively on individuals with annual income above 130,587, and shows that their debt-to-income ratio remains relatively high. This indicates that affluent individuals often carry substantial financial obligations - such as large investments or leveraged assets - which can sustain or even elevate their risk score despite their earnings.

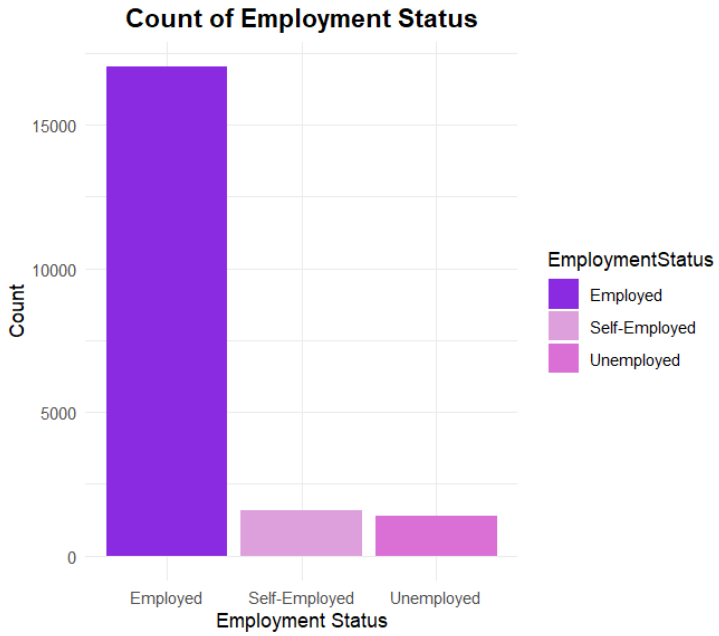


4 Employment Status

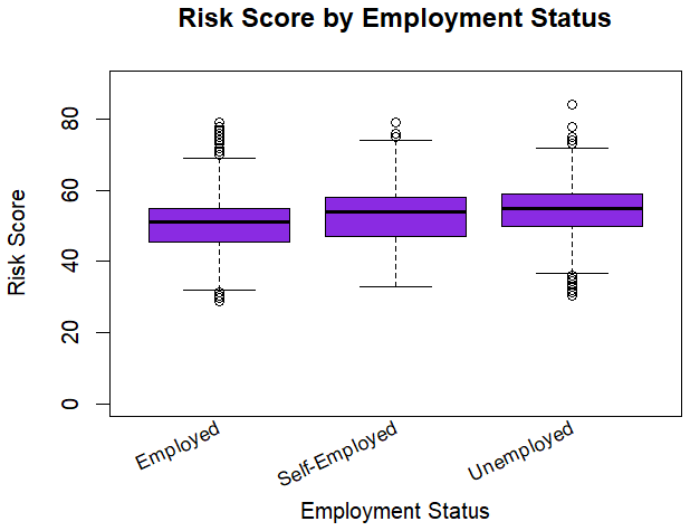
In this chapter, we examine whether an applicant's employment status influences their risk score. In particular, we assmue that there may be differences in risk score outcomes between applicants which are employed compare to those who are self-employed and unemployed.

The counts for each category are as follows:

- Employed: 17036
- Self-Employed: 1573
- Unemployed: 1391



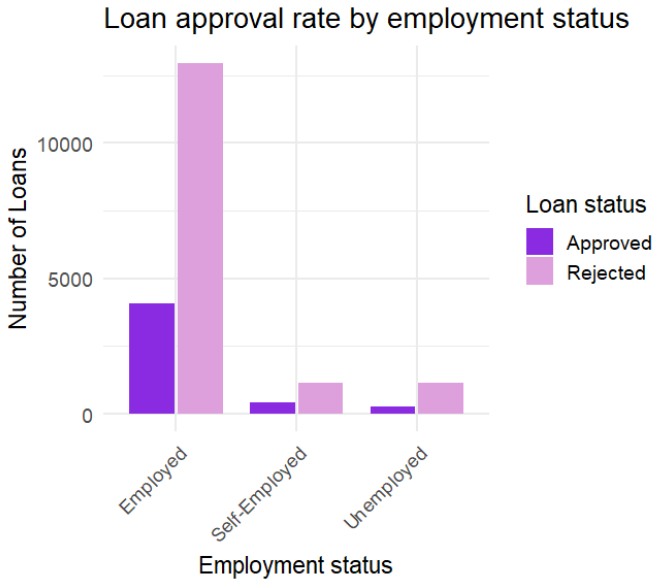
The boxplots for each group of employment status:

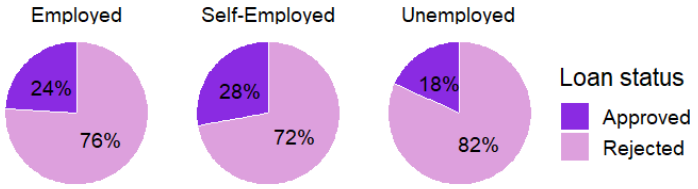


The median allows us to draw conclusions about the performance of at least half of the applicants within each employment category. Differences in these medians across groups highlight how risk scores vary between employed, self-employed and unemployed applicants. The table below presents the mean risk scores of applicants grouped by their employment status.

Employment Status	Employed	Self-Employed	Unemployed
Median	51	54	55

In addition, we present the loan approval status:

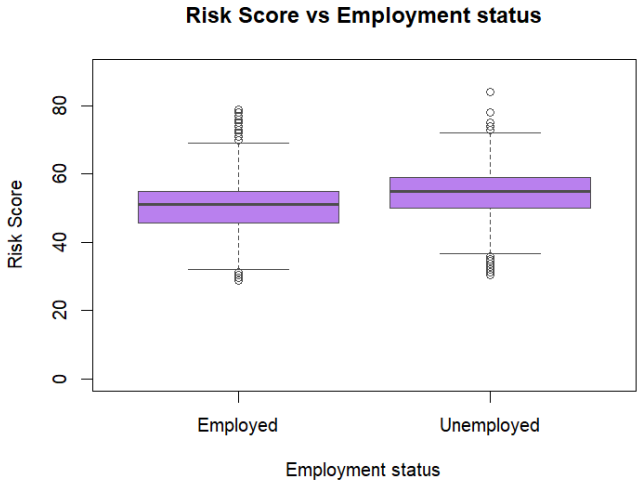




When employment status is taken into account, the share of approved loan applications is relatively similar across groups: roughly 20–25% of applications are approved, while the remaining majority are rejected. Indicating that employment status does not substantially change the overall approval rate.

4.1 Impact of Employment Status on Risk Score

The objective of this analysis was to examine whether applicants who are employed (employed, self-employed) show different risk scores compared to those who are unemployed.
To present and compare the differences in risk scores between these two groups, we demonstrate the boxplot:



The boxplot shows a clear separation between the two employment groups. Unemployed applicants tend to have higher risk scores than employed applicants, with the median for the unemployed group visibly above the median for the employed group.

- Mean risk score for the employed group: $\mu_{\text{emp}} = 50.52902$
- Mean risk score for the unemployed group: $\mu_{\text{unemp}} = 53.94752$
- Overall mean across all applicants: $\mu_{\text{total}} = 50.76678$

4.1.1 Hypothesis Formulation

Let μ_{emp} and μ_{unemp} represent the mean risk scores for the respective groups.

- **Null hypothesis** H_0:

  $$\mu_{\text{emp}} = \mu_{\text{unemp}}$$

  There is no statistically significant difference in average risk scores between applicants who are employed and unemployed.

- **Alternative hypothesis** H_1:

  $$\mu_{\text{emp}} \neq \mu_{\text{unemp}}$$

  The average risk scores differ significantly depending on employment status.

In the context of a linear regression model, this comparison can be reframed as a test of coefficient equality:

$$H_0 : \beta_1 = \beta_2 = \beta_{12} \qquad \text{vs.} \qquad H_1 : \beta_1 \neq \beta_2$$

Here, β_1 and β_2 denote the estimated effects of each employment status on the predicted risk score.

Full Model, defined by the formula:

$$\mu(R_{i1}, R_{i2}) = \beta_1 \cdot R_{i1} + \beta_2 \cdot R_{i2}$$

where:

$$R_{i1} = \begin{cases} 1 & \text{if the applicant are employed} \\ 0 & \text{otherwise} \end{cases}$$

and

$$R_{i2} = \begin{cases} 1 & \text{if the applicant are unemployed} \\ 0 & \text{otherwise} \end{cases}$$

Result:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       50.52902    0.05666  891.75   <2e-16 ***
EmpGroupUnemployed  3.41850    0.21486   15.91   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.73 on 19998 degrees of freedom
Multiple R-squared:  0.0125,    Adjusted R-squared:  0.01245
F-statistic: 253.1 on 1 and 19998 DF,  p-value: < 2.2e-16
```

Reduced Model, defined as:

$$\mu(R_{i1}, R_{i2}) = \beta_{12} \cdot (R_{i1} + R_{i2}) = \beta_{12}$$

Result:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   50.767      0.055     923   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.778 on 19999 degrees of freedom
```

4.1.2 Statistical Testing

We performed an F-test to compare the fit of the full model to teh reduced model. The F-statistic is calculated as:

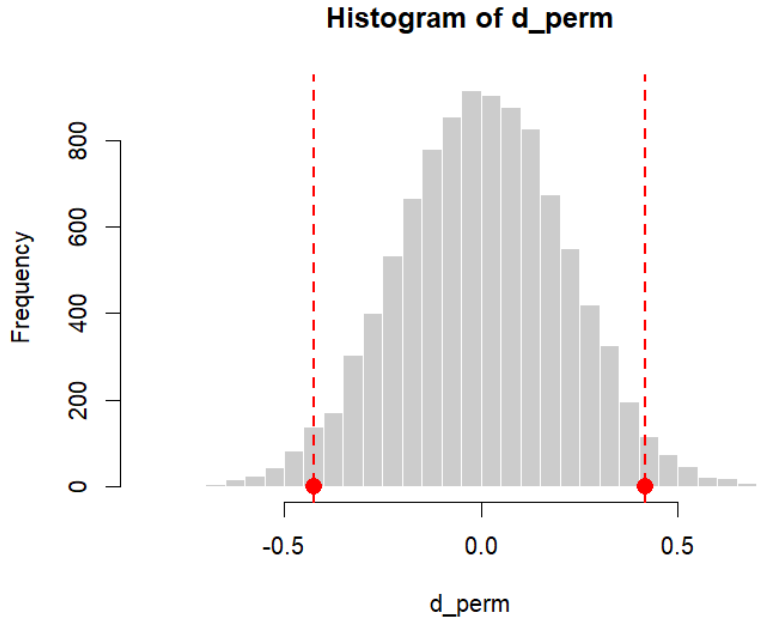
$$F_{\text{obs}} = \frac{\text{SSE}_0 - \text{SSE}_1}{df_0 - df_1} \cdot \frac{df_1}{\text{SSE}_1}$$

In our analysis:

$$F_{\text{obs}} = 253.1433, \qquad F_{\text{crit}} = 3.841924 \quad (\alpha = 0.05)$$

Since the computed F-statistic is larger than the critical value, we reject the null hypothesis and conclude that the full model provides a significantly better fit to the data than the reduced model.

We also performed a Permutation test to assess the differences in risk score between applicants who are employed and unemployed. The observed difference in means is compared to the distribution of differences generated from random permutations



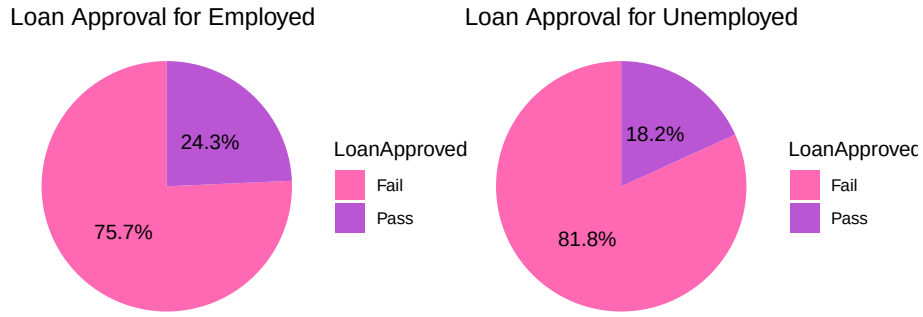
The critical values for rejecting the null hypothesis were obtained from the 2.5th and 97.5th percentiles of the permutation distribution: $L = -0.4238837$ and $U = 0.4155252$. These bounds represent the range of differences in mean risk scores that we would expect to observe under the null hypothesis of no effect of employment status. However, the observed difference in mean risk scores between employed and unemployed applicants ($d_{obs} = -3.418496$) lies far outside this interval. In fact, none of the 10,000 permutations produced a difference as extreme as the observed one, yielding an estimated permutation p-value of $p < 0.0001$. Therefore, we reject the null hypothesis and conclude that employment status is significantly associated with risk score, with unemployed applicants exhibiting substantially higher average risk scores than employed applicants.

Interpretation

Unemployed applicants generally have higher risk scores than employed applicants.

5 Logistic Regression for Employment Status

To examine whether employment status and loan amount influence the likelihood of loan approval, I will apply a logistic regression model. This approach allows us to assess how these predictors affect the probability of receiving a loan.



We consider a binary response variable Y_i indicating whether applicant i was granted a loan:

$$Y_i = \begin{cases} 1 & \text{if the loan was approved} \\ 0 & \text{if the loan was rejected} \end{cases} \quad \text{with} \quad Y_i \mid x_i \sim \text{Bernoulli}(\pi_i)$$

To incorporate employment status, we define indicator variables:

$$R_{i1} = \mathbb{I}(\text{applicant } i \text{ is employed}), \quad R_{i2} = \mathbb{I}(\text{applicant } i \text{ is unemployed})$$

Let x_i denote the loan amount of applicant i. We model the probability of loan approval as a logistic function:

$$\pi_i = \Pr(Y_i = 1 \mid R_{i1}, R_{i2}, x_i) = \frac{\exp\left(\beta_0^E R_{i1} + \beta_1^E R_{i1} x_i + \beta_0^U R_{i2} + \beta_1^U R_{i2} x_i\right)}{1 + \exp\left(\beta_0^E R_{i1} + \beta_1^E R_{i1} x_i + \beta_0^U R_{i2} + \beta_1^U R_{i2} x_i\right)}$$

- β_0^E, β_0^U: intercept terms for employed and unemployed applicants
- β_1^E, β_1^U: slope coefficients for employed and unemployed applicants

This formulation allows both the baseline approval likelihood and the loan amount effect to vary depending on employment status, enabling a flexible comparison of financial behavior across groups.

Let $\beta_0^E, \beta_1^E, \beta_0^U, \beta_1^U$ denote the model parameters for employment and unemployment applicants, respectively. We define the likelihood function for estimating these coefficients based on observed loan decisions:

$$L(\beta_0^E, \beta_1^E, \beta_0^U, \beta_1^U) = \prod_{i \in \mathcal{E}} \left[\pi_i^E(x_i)^{y_i}(1-\pi_i^E(x_i))^{1-y_i}\right] \cdot \prod_{i \in \mathcal{U}} \left[\pi_i^U(x_i)^{y_i}(1-\pi_i^U(x_i))^{1-y_i}\right]$$

where $\mathcal{E}$ and $\mathcal{U}$ denote the sets of employment and unemployment applicants, $y_i \in \{0, 1\}$ indicates whether applicant i was approved for a loan, x_i is the annual income of applicant i, $\pi_i^E(x_i)$ and $\pi_i^U(x_i)$ are the modeled probabilities of approval for each group.

The odds of loan approval for applicant i are given by:

$$\frac{\pi_i}{1-\pi_i} = \exp\left(\beta_0^E R_{i1} + \beta_1^E R_{i1} x_i + \beta_0^U R_{i2} + \beta_1^U R_{i2} x_i\right)$$

and the corresponding log-odds formulation becomes:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0^E R_{i1} + \beta_1^E R_{i1} x_i + \beta_0^U R_{i2} + \beta_1^U R_{i2} x_i$$

For clarity, the model simplifies to group-specific logistic functions:

$$\pi_i^E = \frac{\exp(\beta_0^E + \beta_1^E x_i)}{1 + \exp(\beta_0^E + \beta_1^E x_i)}, \quad \pi_i^U = \frac{\exp(\beta_0^U + \beta_1^U x_i)}{1 + \exp(\beta_0^U + \beta_1^U x_i)}$$

This formulation allows us to estimate separate approval dynamics for employment and unemployment applicants, capturing both baseline differences and loan amount-dependent effects.

5.1 Logistic Regression: Full and Reduced Models

- **Full model** The full model allows both the intercept and the income effect to vary across employment groups:

  $$\gamma(x_i, R_{i1}, R_{i2}) = \beta_0^E R_{i1} + \beta_1^E R_{i1} \cdot x_i + \beta_0^U R_{i2} + \beta_1^U R_{i2} \cdot x_i$$

  where $\gamma(x_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$ is the log-odds of loan approval for applicant i, and x_i denotes their loan amount.

  The corresponding probability of loan approval is given by:

  $$\pi_i = \Pr(Y_i = 1 \mid R_{i1}, R_{i2}, x_i) = \frac{\exp\left(\gamma(x_i, R_{i1}, R_{i2})\right)}{1 + \exp\left(\gamma(x_i, R_{i1}, R_{i2})\right)}$$

- **Reduced model** To test whether the effect of loan amount is consistent across employment groups, we consider a reduced model with a shared slope:

$$\gamma(x_i, R_{i1}, R_{i2}) = \beta_0^E R_{i1} + \beta_0^U R_{i2} + \beta_1^{EU} \cdot x_i$$

This formulation assumes that loan amount influences loan approval similarly for both employment and unemployment applicants, while allowing for group-specific baseline probabilities.

5.2 Hypothesis Testing: Loan Amount Effect by Employment Status

To assess whether the impact of loan amount on the probability of loan approval differs across employment groups, we formulate the following hypotheses:

- **Null Hypothesis (H_0):** The effect of loan amount on approval probability is the same for employed and unemployed applicants:

$$\beta_1^{\mathrm{E}} = \beta_1^{\mathrm{U}}$$

- **Alternative Hypothesis (H_1):** The effect of loan amount differs between the two employment groups:

$$\beta_1^{\mathrm{E}} \neq \beta_1^{\mathrm{U}}$$

To test this, we compare the full and reduced logistic regression models using the Likelihood Ratio Test (LRT). The test statistic is defined as:

$$\mathrm{LRT} = -2 \log \left(\frac{L_0}{L_1} \right)$$

where L_0 is the maximized likelihood under the reduced model (common slope), L_1 is the maximized likelihood under the full model (group-specific slopes).

Conclusions:

```
Analysis of Deviance Table

Model 1: LoanApproved ~ LoanAmount + EmploymentStatus
Model 2: LoanApproved ~ LoanAmount * EmploymentStatus
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     19997      20485
2     19996      20484  1  0.55138   0.4578
```

The LRT statistic is 0.55138, which is far below the chi-square threshold of 3.841 for 1 degree of freedom at the 5% significance level. The associated p-value is 0.4578, which is much greater than 0.05.

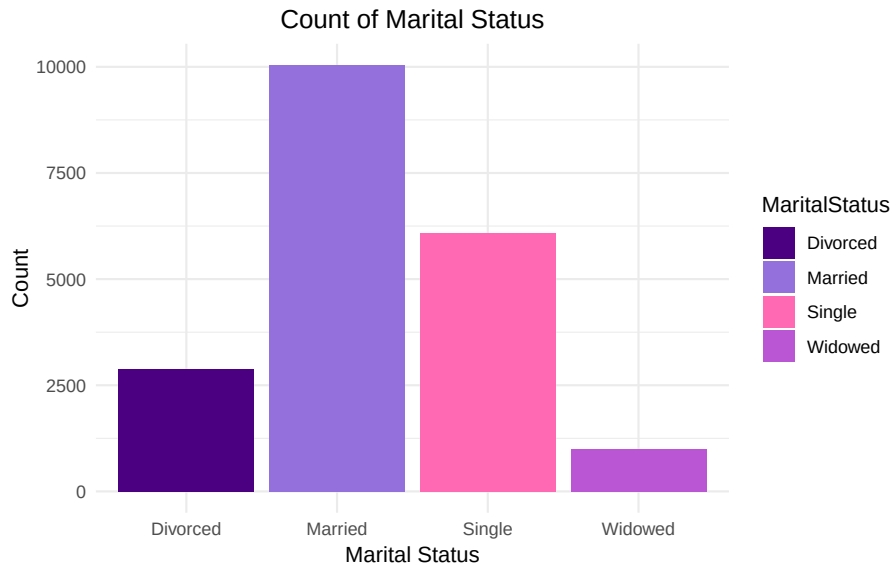
This means that adding the interaction between loan amount and employment status does not make the model noticeably better. In practice, the influence of loan amount on the chance of getting a loan looks very similar for employed and unemployed applicants, so the simpler model without this extra term is enough.

6 Marital Status

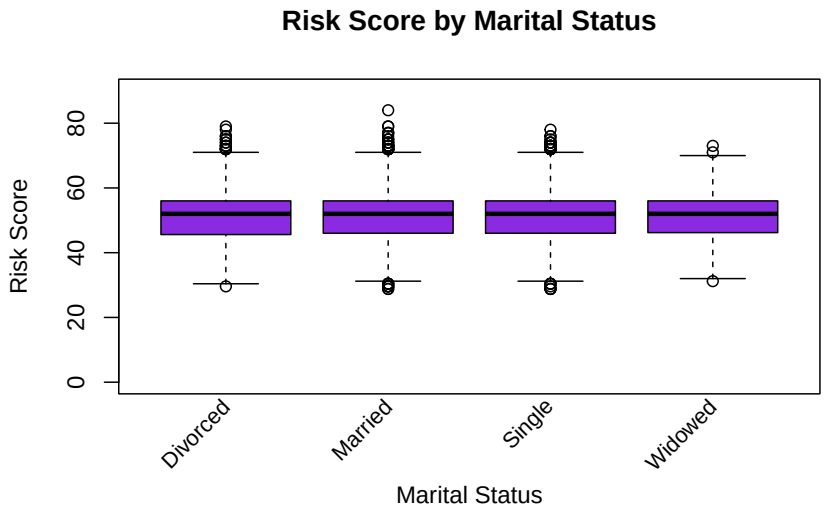
This chapter aims to explore the relationship between marital status and academic performance.

The counts for each category are as follows:

- Divorced: 2882
- Married: 10041
- Single: 6078
- Widowed: 999



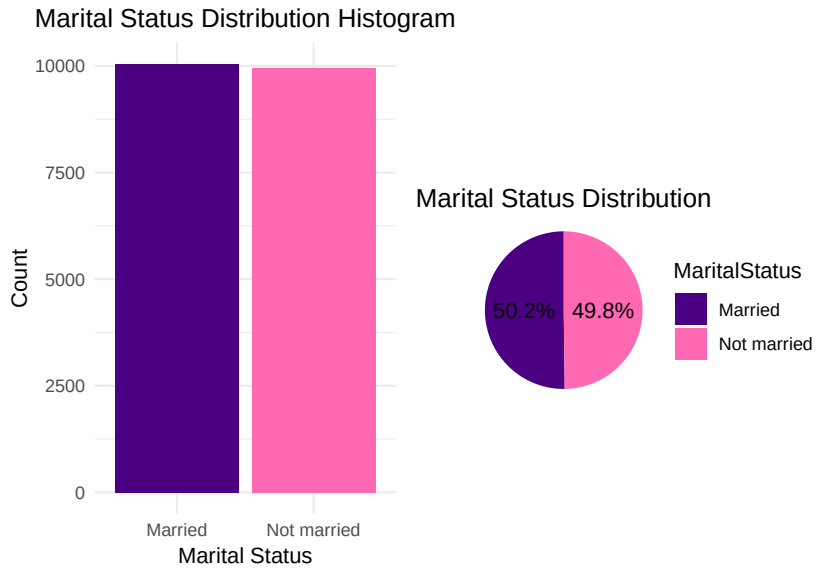
The boxplots for each group of marital status:



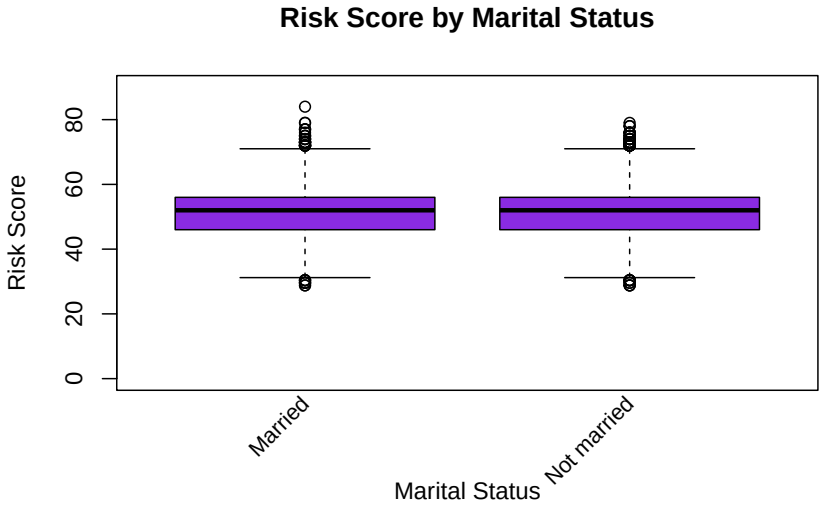
We recategorize the variable Marital Status into two groups: Married, which includes only individuals originally labeled as Married, and Not Married, which combines the categories" Single, Widowed and Divorced.

Now applicants are divided into two categories as follows:

- Married: 10041
- Not married: 9959



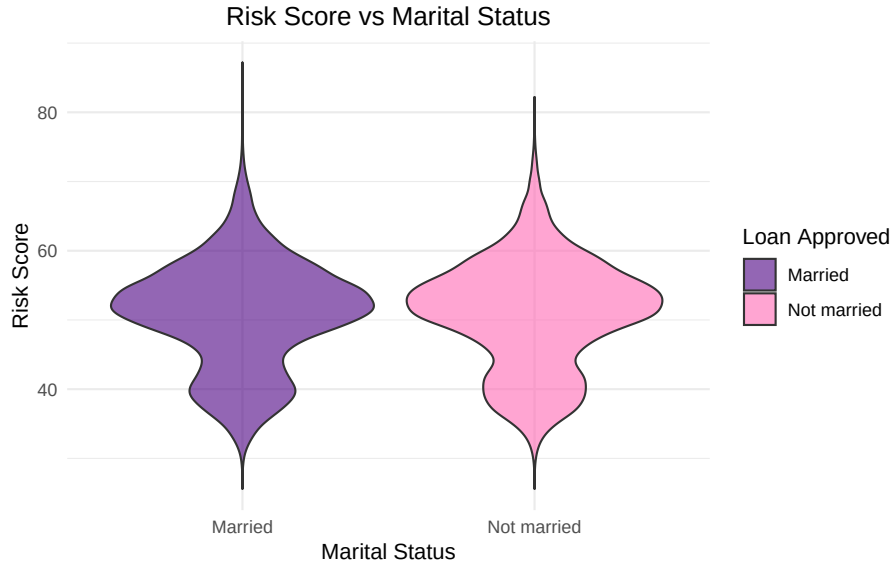
The marital status distribution shows a nearly equal split between married and not married individuals. Both categories are similarly represented, with married accounting for 50.2% and not married for 49.8% of the population.



Although the average risk scores for married and not married individuals are closely aligned, subtle differences in central tendency can still be observed. The mean score for married applicants is slightly lower than that of the not married group, yet both hover around the overall population mean.

- Mean risk score for married individuals: $\mu_{\text{mer}} = 50.73584$
- Mean risk score for not married individual: $\mu_{\text{not}} = 50.79797$
- Overall mean across all applicants: $\mu_{\text{total}} = 50.76678$

The violin plots illustrate the distribution of risk scores across marital status groups, revealing similar central tendencies for both married and not married individuals.



6.1 Hypothesis Formulation

Let μ_{mer} and μ_{not} represent the mean risk scores for the respective groups. Let's test whether there is a statistically significant difference in average risk scores between married and not married applicants.

- **Null hypothesis** H_0:
$$\beta_M = \beta_N = \beta_{MN}$$
- **Alternative hypothesis** H_1:
$$\beta_M \neq \beta_N$$

The models are defined as follows:

Full Model:
$$\mu(R_{iM}, R_{iN}) = \beta_1 \cdot R_{iM} + \beta_2 \cdot R_{iN}$$

where two indicator variables are defined, one for married and one for not married people:

$$R_{iM} = \begin{cases} 1 & \text{if the applicant is married} \\ 0 & \text{if the applicant is not married} \end{cases}$$

$$R_{iN} = \begin{cases} 1 & \text{if the applicant is not married} \\ 0 & \text{if the applicant is married} \end{cases}$$

Result:

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             50.73584    0.07762 653.602   <2e-16 ***
MaritalStatusNot married  0.06213    0.11000   0.565    0.572
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.778 on 19998 degrees of freedom
Multiple R-squared:  1.595e-05, Adjusted R-squared:  -3.405e-05
F-statistic: 0.319 on 1 and 19998 DF,  p-value: 0.5722
```

Reduced Model:

$$\mu(R_{iM}, R_{iN}) = \beta_{MN} \cdot (R_{iM} + R_{iN}) = \beta_{MN}$$

where β_{MN} is common mean of those groups.

Result:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   50.767      0.055     923   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.778 on 19999 degrees of freedom
```

6.2 Statistical Testing

To formally evaluate whether the marital status is associated with significant differences in applicants' risk scores, we apply **F-statistic** and the **t-test**.

The **F-statistic** is computed as:

$$F_{\text{obs}} = \frac{\text{SSE}_0 - \text{SSE}_1}{df_0 - df_1} \cdot \frac{df_1}{\text{SSE}_1}$$

In our analysis:

$$F_{\text{obs}} = 0.3189822, \qquad F_{\text{crit}} = 3.841924 \quad (\alpha = 0.05)$$

Since the observed F-statistic is lower than the critical value, we fail to reject the null hypothesis. This indicates that marital status does not have a statistically significant effect on applicants' average risk scores.

To test whether the average risk scores differ between the two education groups, we also use the **t-test**:

$$t_{\text{obs}} = \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_0}\right)}}$$

We compare the observed t-statistic to the critical values for a two-tailed test at significance level $\alpha = 0.05$:

$$t_{\text{obs}} = 0.564, \qquad t_{\text{left}} = \text{-}1.96, \qquad t_{\text{right}} = 1.96$$

Since $t_{\text{left}} < t_{\text{obs}} < t_{\text{right}}$, we do not reject the null hypothesis, indicating a statistically significant difference in mean risk scores between the two marital status.

Interpretation Statistical tests indicate that marital status does not significantly influence applicants' average risk scores. Both the F-test and t-test results support the conclusion that differences between married and not married groups are not statistically meaningful

6.3 The analysis of impact of Marital Status on Risk Score and Annual Income

To examine whether the relationship between risk score and annual income varies across marital status, so similarly to above we define two binary indicators that distinguish applicants by their status:

$$R_{iM} = \begin{cases} 1 & \text{if the applicant is married} \\ 0 & \text{otherwise} \end{cases}$$

$$R_{iN} = \begin{cases} 1 & \text{if the applicant is not married} \\ 0 & \text{otherwise} \end{cases}$$

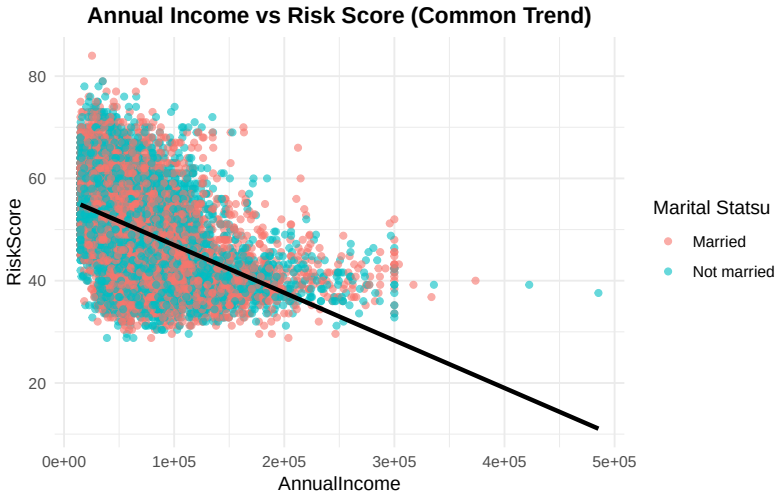
6.3.1 Modeling a Common Trend in Annual Income by Marital Status

First, we specify a simplified model that assumes a shared slope for both groups. Under the hypothesis

$$H_0 : \beta_1^M = \beta_1^N = \beta_1^{MN}$$

we define model:

$$\mu(x_i, R_{iM}, R_{iN}) = \beta_0^1 R_{iM} + \beta_0^2 R_{iN} + \beta_1^{MN} \cdot x_i$$



6.3.2 Modeling a Different Trends in Interest Rate by Education Level

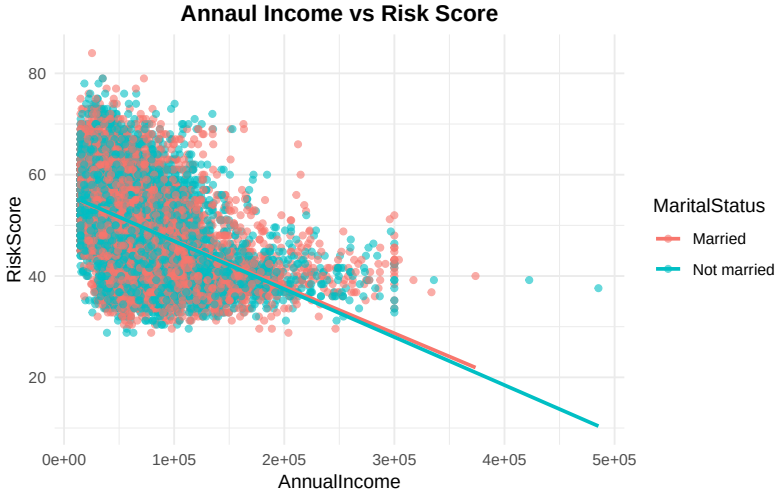
Assuming that the effect of marital status on risk score may follow different trends in each group, under the hypothesis

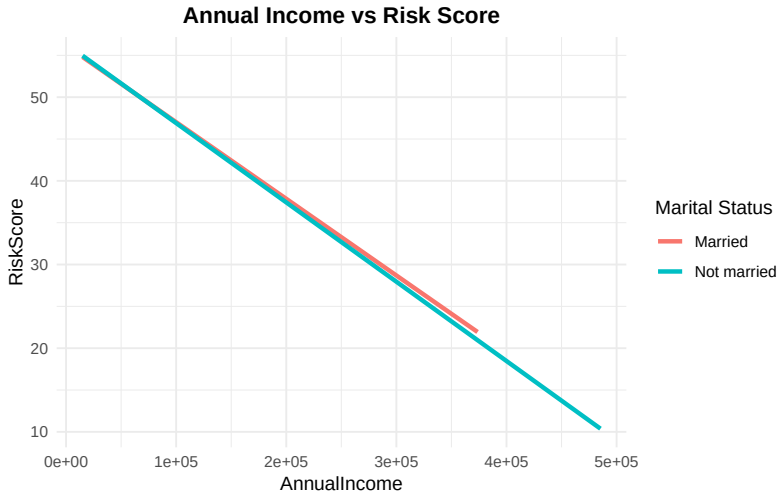
$$H_1 : \beta_1^M \neq \beta_1^N$$

we specify the full model as:

$$\mu(x_i, R_{iM}, R_{iN}) = \beta_0^M R_{iM} + \beta_1^M R_{iM} \cdot x_i + \beta_0^N R_{iN} + \beta_1^N R_{iN} \cdot x_i$$

where





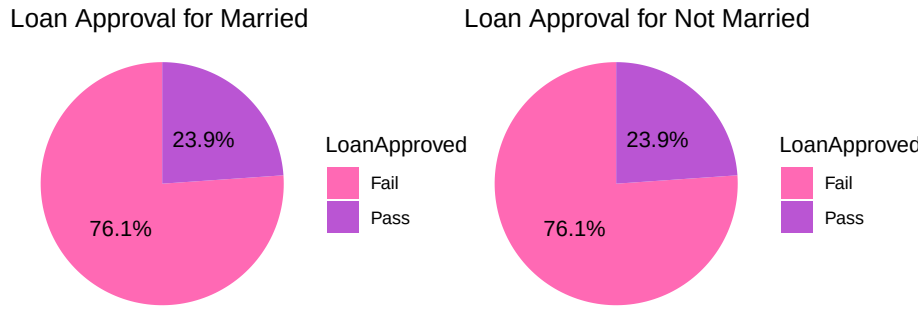
6.3.3 Conclusions

Since the observed F-statistic (1.9401) is lower than the critical value (3.84), we do not reject the null hypothesis. This suggests that the interaction between risk score and marital status does not significantly improve the model's ability to explain annual income.

```
Analysis of Variance Table

Model 1: AnnualIncome ~ RiskScore + MaritalStatus
Model 2: AnnualIncome ~ RiskScore * MaritalStatus
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1  19997 2.4956e+13
2  19996 2.4954e+13  1 2421095232 1.9401 0.1637
```

7 Logistic Regression for Marital Status

To examine whether marital status and annual income influence the likelihood of loan approval, I will apply a logistic regression model. This approach allows us to assess how these predictors affect the probability of receiving a loan.



We consider a binary response variable Y_i indicating whether applicant i was granted a loan:

$$Y_i = \begin{cases} 1 & \text{if the loan was approved} \\ 0 & \text{if the loan was rejected} \end{cases} \quad \text{with} \quad Y_i \mid x_i \sim \text{Bernoulli}(\pi_i)$$

To incorporate marital status, we define indicator variables:

$$R_{iM} = \mathbb{I}(\text{applicant } i \text{ is married}), \quad R_{iN} = \mathbb{I}(\text{applicant } i \text{ is not married})$$

Let x_i denote the annual income of applicant i. We model the probability of loan approval as a logistic function:

$$\pi_i = \Pr(Y_i = 1 \mid R_{iM}, R_{iN}, x_i) = \frac{\exp\left(\beta_0^M R_{iM} + \beta_1^M R_{iM} x_i + \beta_0^N R_{iN} + \beta_1^N R_{iN} x_i\right)}{1 + \exp\left(\beta_0^M R_{iM} + \beta_1^M R_{iM} x_i + \beta_0^N R_{iN} + \beta_1^N R_{iN} x_i\right)}$$

Let $\beta_0^M, \beta_1^M, \beta_0^N, \beta_1^N$ denote the model parameters for married and not married applicants, respectively. We define the likelihood function:

$$L(\beta_0^M, \beta_1^M, \beta_0^N, \beta_1^N) = \prod_{i \in \mathcal{M}} \left[\pi_i^M(x_i)^{y_i}(1 - \pi_i^M(x_i))^{1-y_i}\right] \cdot \prod_{i \in \mathcal{N}} \left[\pi_i^N(x_i)^{y_i}(1 - \pi_i^N(x_i))^{1-y_i}\right]$$

where $\mathcal{M}$ and $\mathcal{N}$ denote the sets of married and not married applicants, $y_i \in \{0, 1\}$ indicates whether applicant i was approved for a loan, x_i is the annual income

of applicant i, $\pi_i^M(x_i)$ and $\pi_i^N(x_i)$ are the modeled probabilities of approval for each group.

Group-specific logistic functions:

$$\pi_i^M = \frac{\exp(\beta_0^M + \beta_1^M x_i)}{1 + \exp(\beta_0^M + \beta_1^M x_i)}, \quad \pi_i^N = \frac{\exp(\beta_0^N + \beta_1^N x_i)}{1 + \exp(\beta_0^N + \beta_1^N x_i)}$$

7.1 Logistic Regression: Full and Reduced Models

- **Full model** The full model allows both the intercept and the income effect to vary across marital groups:

  $$\gamma(x_i, R_{iM}, R_{iN}) = \beta_0^M R_{iM} + \beta_1^M R_{iM} \cdot x_i + \beta_0^N R_{iN} + \beta_1^N R_{iN} \cdot x_i$$

  where $\gamma(x_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$.

  The probability of loan approval is given by:

  $$\pi_i = \Pr(Y_i = 1 \mid R_{iM}, R_{iN}, x_i) = \frac{\exp\left(\gamma(x_i, R_{iM}, R_{iN})\right)}{1 + \exp\left(\gamma(x_i, R_{iM}, R_{iN})\right)}$$

- **Reduced model** To test whether the effect of income is consistent across marital groups, we consider a reduced model with a shared slope:

  $$\gamma(x_i, R_{iM}, R_{iN}) = \beta_0^M R_{iM} + \beta_0^N R_{iN} + \beta_1^{MN} \cdot x_i$$

7.2 Hypothesis Testing: Annual Income Effect by Marital Status

To assess whether the impact of annual income on the probability of loan approval differs across groups, we formulate the following hypotheses:

- **Null Hypothesis (H_0):** The effect of on approval probability is the same for both groups:

  $$\beta_1^{\mathrm{M}} = \beta_1^{\mathrm{N}}$$

- **Alternative Hypothesis (H_1):** The effect of loan amount differs between the groups:

  $$\beta_1^{\mathrm{M}} \neq \beta_1^{\mathrm{N}}$$

To test this, we compare the full and reduced logistic regression models using the Likelihood Ratio Test (LRT):

$$\text{LRT} = -2 \log \left(\frac{L_0}{L_1} \right).$$

Conclusions:

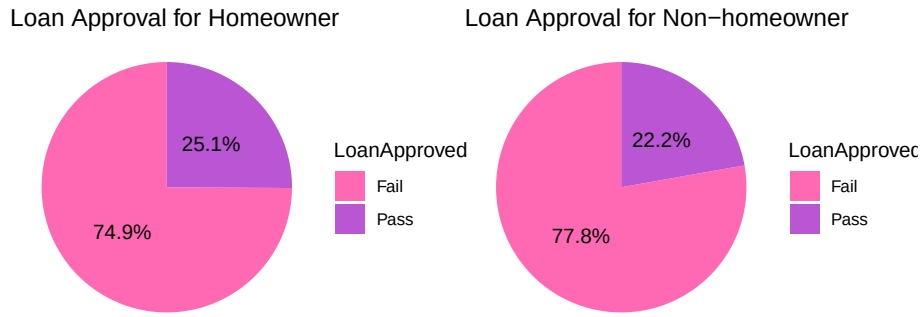
```
Analysis of Deviance Table

Model 1: LoanApproved ~ AnnualIncome + MaritalStatus
Model 2: LoanApproved ~ AnnualIncome * MaritalStatus
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     19997      14285
2     19996      14285  1  0.13841   0.7099
```

The LRT statistic is 0.13841, which is far below the chi-square threshold of 3.841 for 1 degree of freedom at the 5% significance level. The associated p-value is 0.7099, which is much greater than 0.05.

These results indicate that there is no statistical evidence to suggest that including the interaction between annual income and marital status improves the model. In other words, we have no basis to assume that the effect of income on loan approval differs meaningfully across marital groups.

8 Logistic Regression for Home Ownership Status

To examine whether house ownership and annual income influence the likelihood of loan approval, I will apply a logistic regression model. This approach allows us to assess how these predictors affect the probability of receiving a loan.



We consider a binary response variable Y_i indicating whether applicant i was granted a loan:

$$Y_i = \begin{cases} 1 & \text{if the loan was approved} \\ 0 & \text{if the loan was rejected} \end{cases} \quad \text{with} \quad Y_i \mid x_i \sim \text{Bernoulli}(\pi_i)$$

To incorporate marital status, we define indicator variables:

$$R_{iH} = \mathbb{I}(\text{applicant } i \text{ own a house}), \quad R_{iN} = \mathbb{I}(\text{applicant } i \text{ does not own a house})$$

Let x_i denote the annual income of applicant i. We model the probability of loan approval as a logistic function:

$$\pi_i = \Pr(Y_i = 1 \mid R_{iH}, R_{iN}, x_i) = \frac{\exp\left(\beta_0^H R_{iH} + \beta_1^H R_{iH} x_i + \beta_0^N R_{iN} + \beta_1^N R_{iN} x_i\right)}{1 + \exp\left(\beta_0^H R_{iH} + \beta_1^H R_{iH} x_i + \beta_0^N R_{iN} + \beta_1^N R_{iN} x_i\right)}$$

Let $\beta_0^H, \beta_1^H, \beta_0^N, \beta_1^N$ denote the model parameters for homeowner and non-homeowner applicants, respectively. We define the likelihood function:

$$L(\beta_0^H, \beta_1^H, \beta_0^N, \beta_1^N) = \prod_{i \in \mathcal{H}} \left[\pi_i^H(x_i)^{y_i}(1 - \pi_i^H(x_i))^{1-y_i}\right] \cdot \prod_{i \in \mathcal{N}} \left[\pi_i^N(x_i)^{y_i}(1 - \pi_i^N(x_i))^{1-y_i}\right]$$

where $\mathcal{H}$ and $\mathcal{N}$ denote the sets of homeowner and non-homeowner applicants, $y_i \in \{0, 1\}$ indicates whether applicant i was approved for a loan, x_i is the annual income of applicant i, $\pi_i^H(x_i)$ and $\pi_i^N(x_i)$ are the modeled probabilities of approval for each group.

Group-specific logistic functions:

$$\pi_i^H = \frac{\exp(\beta_0^H + \beta_1^H x_i)}{1 + \exp(\beta_0^H + \beta_1^H x_i)}, \quad \pi_i^N = \frac{\exp(\beta_0^N + \beta_1^N x_i)}{1 + \exp(\beta_0^N + \beta_1^N x_i)}$$

8.1 Logistic Regression: Full and Reduced Models

- **Full model** The full model allows both the intercept and the income effect to vary across ownership groups:

  $$\gamma(x_i, R_{iH}, R_{iN}) = \beta_0^H R_{iH} + \beta_1^H R_{iH} \cdot x_i + \beta_0^N R_{iN} + \beta_1^N R_{iN} \cdot x_i$$

  where $\gamma(x_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$.

  The probability of loan approval is given by:

  $$\pi_i = \Pr(Y_i = 1 \mid R_{iH}, R_{iN}, x_i) = \frac{\exp\left(\gamma(x_i, R_{iH}, R_{iN})\right)}{1 + \exp\left(\gamma(x_i, R_{iH}, R_{iN})\right)}$$

- **Reduced model** To test whether the effect of income is consistent across ownership groups, we consider a reduced model with a shared slope:

  $$\gamma(x_i, R_{iH}, R_{iN}) = \beta_0^H R_{iH} + \beta_0^N R_{iN} + \beta_1^{HN} \cdot x_i$$

8.2 Hypothesis Testing: Annual Income Effect by Marital Status

To assess whether the impact of annual income on the probability of loan approval differs across groups, we formulate the following hypotheses:

- **Null Hypothesis (H_0):** The effect of on approval probability is the same for both groups:

  $$\beta_1^{\mathrm{H}} = \beta_1^{\mathrm{N}}$$

- **Alternative Hypothesis (H_1):** The effect of loan amount differs between the groups:

  $$\beta_1^{\mathrm{H}} \neq \beta_1^{\mathrm{N}}$$

To test this, we compare the full and reduced logistic regression models using the Likelihood Ratio Test (LRT):

$$\text{LRT} = -2 \log \left(\frac{L_0}{L_1} \right).$$

Conclusions:

```
Analysis of Deviance Table

Model 1: LoanApproved ~ AnnualIncome + HomeOwnershipStatus
Model 2: LoanApproved ~ AnnualIncome * HomeOwnershipStatus
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     19997      14240
2     19996      14238  1   2.1618   0.1415
```

The LRT statistic is 2.1618, which is below the chi-square threshold of 3.841 for 1 degree of freedom at the 5% significance level. The associated p-value is 0.1415, which is greater than 0.05.

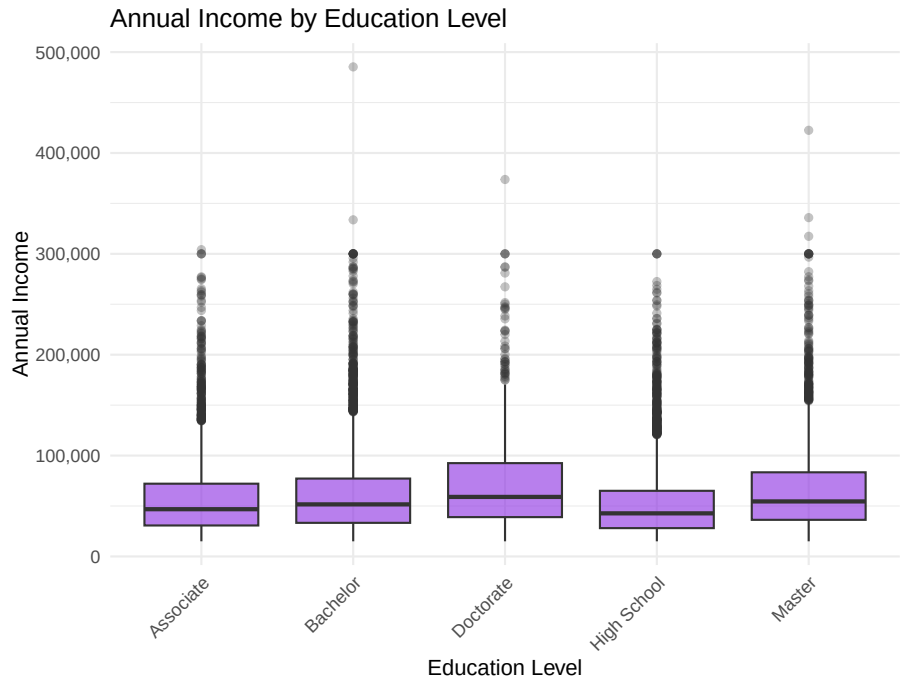
These results indicate that there is no statistical evidence to suggest that including the interaction between annual income and home ownership status improves the model. In other words, we have no basis to assume that the effect of income on loan approval differs meaningfully across ownership groups.

9 Exploratory Analysis of Predictor Relationships

In this section, we explore relationships between key applicant characteristics, independently of the outcome variables

9.1 Annual Income vs Education Level

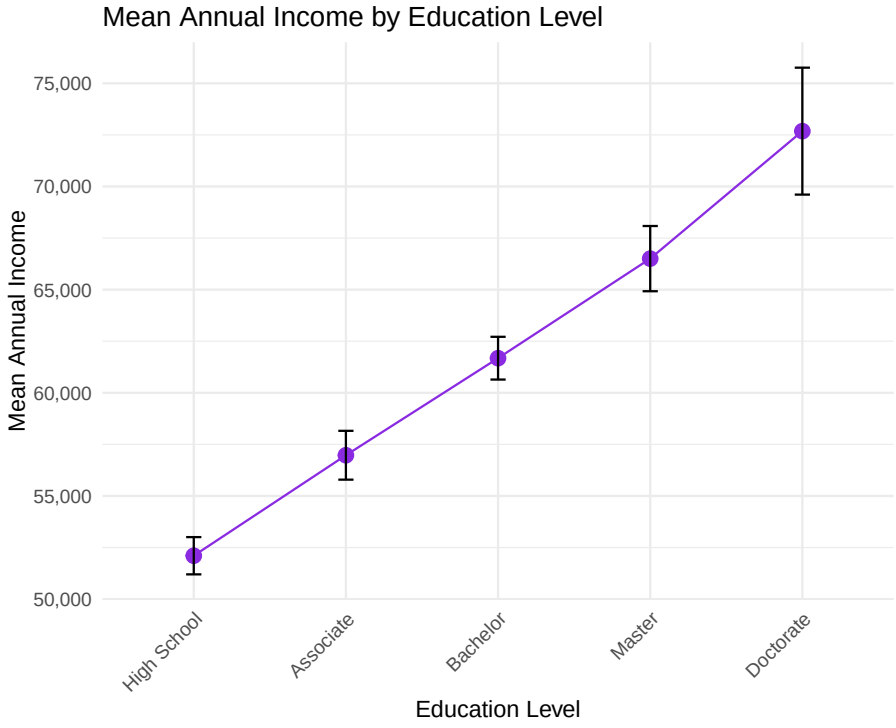
In this subsection, we investigate the relationship between applicants' education level and their annual income.



The boxplots show a clear upward trend in annual income across education levels. Applicants with higher education (Bachelor, Master, Doctorate) tend to earn more on average than those with only a high school or associate degree, as indicated by the increasing medians and upper quartiles. At the same time, the distributions are quite wide and strongly overlapping, suggesting substantial income variability within each education group.

Education level	Mean Income	Median Income	First quartile	Third quartile
High School	52,103	42,722	28,080	65,070
Associate	56,973	46,834	30,698	72,129
Bachelor	61,678	51,626	33,308	77,204
Master	66,503	54,544	36,305	83,418
Doctorate	72,683	59,025	38,981	92,512

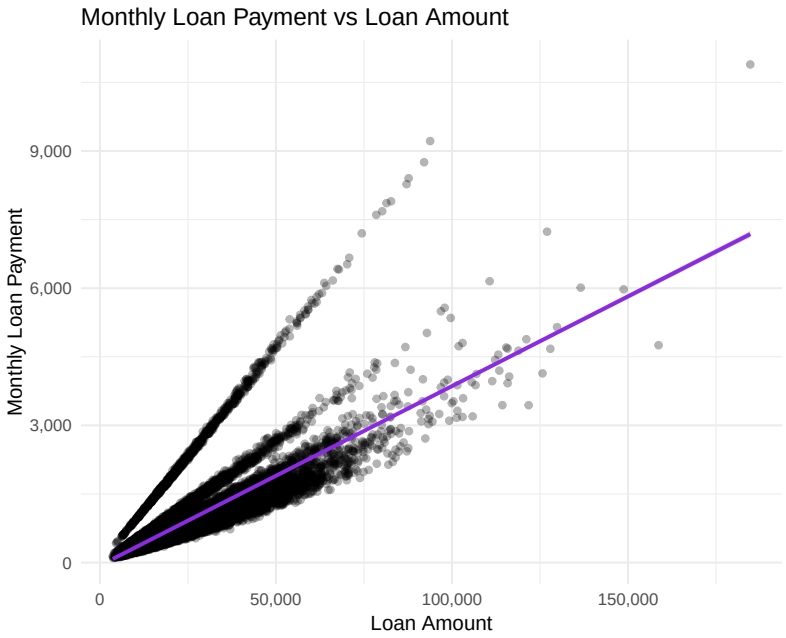
The summary statistics in table confirm the pattern observed in the boxplots. Both the mean and median annual income increase steadily with education level: applicants with a doctorate earn on average about $20,000$ more per year than those with only a high school diploma. In addition, the first and third quartiles also shift upward across education groups, indicating that not only the typical income but the entire income distribution moves to higher levels as education increases.



The mean annual income across education levels, together with 95% confidence intervals, shows an almost linear increase from high school to doctorate. The confidence bands are relatively narrow, indicating that this upward trend is both strong and precisely estimated.

9.2 Monthy Loan Payment vs Loan Amount

In this subsection, we investigate the relationship between the Monthly Loan Payment and the Loan Amount. Our goal is to assess how strongly the loan size determines the required monthly installment and whether this relationship appears approximately linear. We also look for patterns that may suggest the influence of other factors, such as loan duration or interest rate, on this dependence.



The plot shows a strong positive association: higher loan amounts are generally linked to higher monthly payments, although the spread around the line suggests that other factors, such as loan duration and interest rate, also affect the exact payment level.

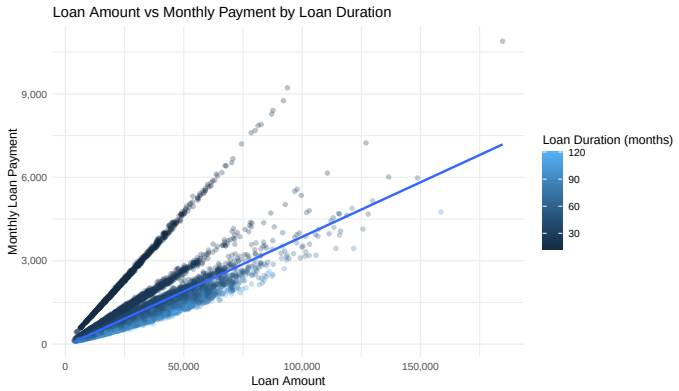
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.450e+01  6.275e+00  -10.28   <2e-16 ***
LoanAmount   3.923e-02  2.219e-04  176.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 421.5 on 19998 degrees of freedom
Multiple R-squared:  0.6097,    Adjusted R-squared:  0.6097
F-statistic: 3.124e+04 on 1 and 19998 DF,  p-value: < 2.2e-16
```

A simple linear regression model was fitted with Monthly Loan Payment as the response and Loan Amount as the predictor. The estimated regression equation is Monthly Loan Payment $= -64.5 + 0.0392 \cdot$ Loan Amount. The slope coefficient is highly statistically significant ($p < 0.001$) and implies that an increase in the loan amount by 1,000 monetary units is associated with an average increase in the monthly payment of about 39 units. The model explains approximately 61 % of the variance in monthly payments ($R^2 = 0.61$), indicating a strong positive linear relationship between loan size and required monthly installment, although the negative intercept has no meaningful financial interpretation for a loan amount equal to zero

9.2.1 Monthly Loan Payment vs Loan Amount by Loan Duration

In this subsection, we examine how the relationship between Monthly Loan Payment and Loan Amount changes across different Loan Duration values. By comparing

these groups, we aim to see whether longer repayment periods systematically reduce the monthly burden for the same loan size.



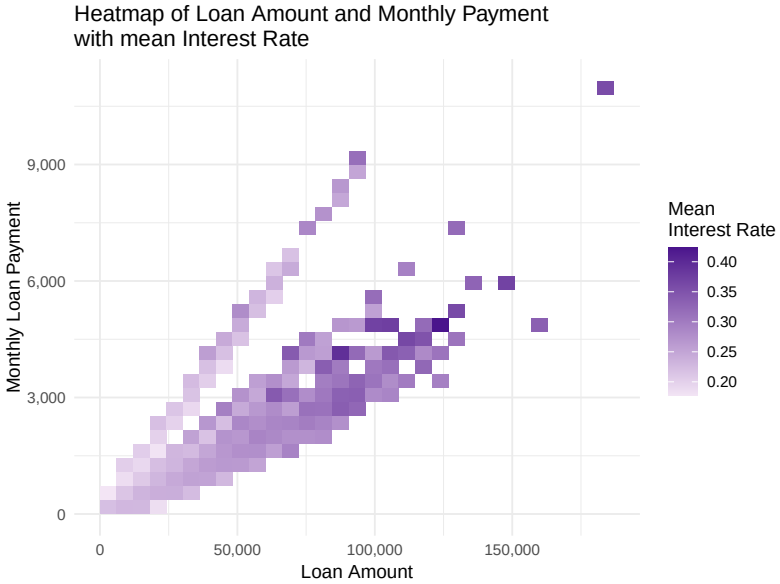
The plot shows that, for a given loan amount, shorter loan durations (darker points) are associated with substantially higher monthly payments, while longer durations (lighter points) spread the same loan amount over lower monthly installments.

```
Correlation tests
Call:r.test(n = nrow(short), r12 = r_short, r34 = r_long, n2 = nrow(long))
Test of difference between two independent correlations
 z value 67.88    with probability  0
```

The correlation between Loan Amount and Monthly Loan Payment equals r = r_{short} for short-term loans and r = r_{long} for long-term loans. A Fisher z-test comparing these correlations yields z = 67.88, p $<$ 0.001, indicating that the strength of the relationship differs significantly between short and long loan durations.

9.2.2 Monthly Loan Payment vs Loan Amount by Interest Rate

In this subsection, we analyse how the relationship between Monthly Loan Payment and Loan Amount varies across different Interest Rate levels.



Darker tiles correspond to higher average interest rates and are concentrated in areas with both higher loan amounts and higher monthly payments, indicating that, for a given loan size, contracts with larger instalments tend to carry higher interest rates.

In the next step, we investigate whether the interest rate has an additional effect on the Monthly Loan Payment beyond the effect of the Loan Amount. Specifically, we compute the partial correlation between Monthly Loan Payment and Interest Rate while controlling for Loan Amount, to assess the association between these two variables for a fixed loan size.

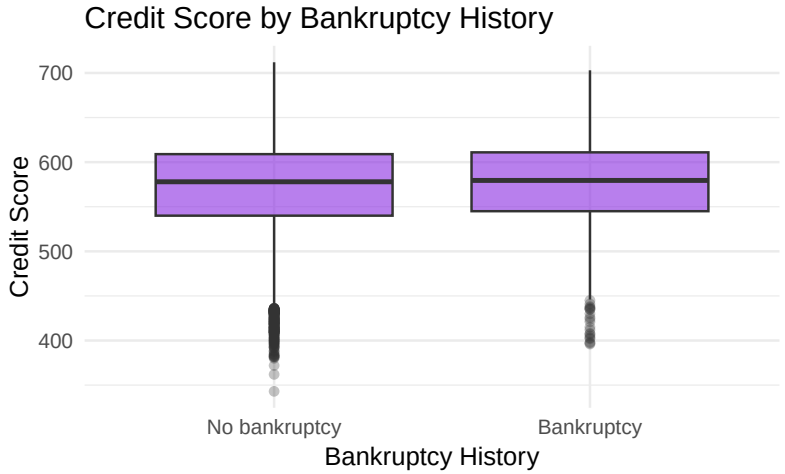
```
    estimate       p.value statistic     n gp  Method
1 -0.2106771 1.937939e-199 -30.47601 20000  1 pearson
```

The partial correlation between Monthly Loan Payment and Interest Rate controlling for Loan Amount equals r = -0.21 ($p < 0.001$, n = 20,000). This negative and statistically significant association indicates that, for loans of similar size, higher interest rates tend to be linked with slightly lower monthly instalments, which is consistent with the idea that higher-rate loans are often repaid over longer durations, reducing the monthly payment despite the higher rate.

9.3 CreditScore vs BankruptcyHistory

In this subsection, we examine how applicants' credit scores differ depending on whether they have a prior history of bankruptcy, they are divided into to categorities as follows:

- No bankruptcy: 18,952
- Bankrupty: 1,048



The boxplot shows that credit scores are fairly similar for applicants with and without a history of bankruptcy, with only a small difference in the central tendency between the two groups.

- Mean credit score for No bankruptcy: 572
- Mean credit score for bankruptcy: 573
- Overall mean across all applicants: 572

To formally examine whether the average credit score differs between borrowers with and without a history of bankruptcy, we apply a Welch two-sample t-test, which compares the group means while allowing for unequal variances.

```
        Welch Two Sample t-test

data:  CreditScore by BankruptcyHistory
t = -1.215, df = 1164.3, p-value = 0.2246
alternative hypothesis: true difference in means between group No bankruptcy and group Bankruptcy is not equal to
0
95 percent confidence interval:
 -5.169233  1.215482
sample estimates:
mean in group No bankruptcy    mean in group Bankruptcy
                   571.5088                    573.4857
```

The Welch t-test comparing mean credit scores of borrowers with and without a bankruptcy record yields t(1164.3) = -1.21, p = 0.225. The mean credit score is 571.5 for borrowers without bankruptcy and 573.5 for those with a bankruptcy history, a difference of only about 2 points. The 95% confidence interval for the difference in means ranges from -5.17 to 1.22 and includes zero. Therefore, there is no statistically significant evidence that the average credit score differs between the two groups in this dataset; the small observed difference is consistent with random variation.

10 Variable Selection with respect to Risk Score

In our dataset, many applicants and loan characteristics can affect the RiskScore. To find which variables are the most important, we use variable selection methods (forward, backward and stepwise) on the training data. These methods help us keep only the variables that have a clear and meaningful impact on the risk score.

Using a predictive modelling approach, we split the loan dataset into two parts: a training set and a test set. About 80% of the observations are used for training, while the remaining 20% are kept for testing the model. On the training set we consider two starting models for the RiskScore:

- **Full Model**, which uses all available explanatory variables in the dataset,
- **Null Model**, which includes only an intercept and no explanatory variables.

We applied automatic variable selection procedures on the training set to choose the most suitable models:

10.1 Variable selection on the original dataset

In this subsection we apply forward, backward and stepwise selection to the original training sample in order to identify which variables are most relevant for explaining the RiskScore. Starting from the full and null models defined above, we use the step() function to obtain candidate linear models and compare them using information criteria and prediction errors.

1. **FORWARD SELECTION**

```
Coefficients:
                               Estimate Std. Error  t value Pr(>|t|)
(Intercept)                   5.131e+01  6.216e-01   82.537  < 2e-16 ***
MonthlyIncome                -1.083e-03  8.843e-06 -122.450  < 2e-16 ***
BankruptcyHistory1            1.323e+01  1.274e-01  103.856  < 2e-16 ***
DebtToIncomeRatio             1.534e+01  1.776e-01   86.359  < 2e-16 ***
NetWorth                     -2.007e-05  2.463e-07  -81.487  < 2e-16 ***
PreviousLoanDefaults1         6.804e+00  9.576e-02   71.059  < 2e-16 ***
InterestRate                  2.999e+01  9.782e-01   30.665  < 2e-16 ***
LengthOfCreditHistory        -1.697e-01  3.408e-03  -49.804  < 2e-16 ***
EmploymentStatusSelf-Employed 2.711e+00  1.071e-01   25.305  < 2e-16 ***
EmploymentStatusUnemployed    3.606e+00  1.130e-01   31.918  < 2e-16 ***
CreditCardUtilizationRate     4.895e+00  1.786e-01   27.408  < 2e-16 ***
MonthlyLoanPayment            6.828e-04  7.137e-05    9.567  < 2e-16 ***
EducationLevelBachelor       -4.511e-01  8.221e-02   -5.488 4.13e-08 ***
EducationLevelDoctorate      -1.594e+00  1.463e-01  -10.895  < 2e-16 ***
EducationLevelHigh School     3.662e-01  8.265e-02    4.431 9.44e-06 ***
EducationLevelMaster         -1.178e+00  9.784e-02  -12.040  < 2e-16 ***
Age                          -3.026e-02  2.633e-03  -11.493  < 2e-16 ***
CreditScore                  -9.754e-03  7.868e-04  -12.398  < 2e-16 ***
MonthlyDebtPayments           1.168e-03  1.187e-04    9.837  < 2e-16 ***
PaymentHistory               -3.355e-02  5.787e-03   -5.796 6.90e-09 ***
LoanAmount                    2.180e-05  3.920e-06    5.561 2.72e-08 ***
TotalLiabilities              2.659e-06  6.086e-07    4.369 1.26e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.606 on 15978 degrees of freedom
Multiple R-squared:  0.7868,    Adjusted R-squared:  0.7865
F-statistic:  2808 on 21 and 15978 DF,  p-value: < 2.2e-16
```

$$\text{BIC} = 86653.1 \qquad \text{AIC} = 86476.45$$

2. BACKWARD SELECTION

```
Coefficients:
                               Estimate Std. Error  t value Pr(>|t|)
(Intercept)                   5.131e+01  6.216e-01   82.537  < 2e-16 ***
Age                          -3.026e-02  2.633e-03  -11.493  < 2e-16 ***
CreditScore                  -9.754e-03  7.868e-04  -12.398  < 2e-16 ***
EmploymentStatusSelf-Employed 2.711e+00  1.071e-01   25.305  < 2e-16 ***
EmploymentStatusUnemployed    3.606e+00  1.130e-01   31.918  < 2e-16 ***
EducationLevelBachelor       -4.511e-01  8.221e-02   -5.488 4.13e-08 ***
EducationLevelDoctorate      -1.594e+00  1.463e-01  -10.895  < 2e-16 ***
EducationLevelHigh School     3.662e-01  8.265e-02    4.431 9.44e-06 ***
EducationLevelMaster         -1.178e+00  9.784e-02  -12.040  < 2e-16 ***
LoanAmount                    2.180e-05  3.920e-06    5.561 2.72e-08 ***
MonthlyDebtPayments           1.168e-03  1.187e-04    9.837  < 2e-16 ***
CreditCardUtilizationRate     4.895e+00  1.786e-01   27.408  < 2e-16 ***
DebtToIncomeRatio             1.534e+01  1.776e-01   86.359  < 2e-16 ***
BankruptcyHistory1            1.323e+01  1.274e-01  103.856  < 2e-16 ***
PreviousLoanDefaults1         6.804e+00  9.576e-02   71.059  < 2e-16 ***
PaymentHistory               -3.355e-02  5.787e-03   -5.796 6.90e-09 ***
LengthOfCreditHistory        -1.697e-01  3.408e-03  -49.804  < 2e-16 ***
TotalLiabilities              2.659e-06  6.086e-07    4.369 1.26e-05 ***
MonthlyIncome                -1.083e-03  8.843e-06 -122.450  < 2e-16 ***
NetWorth                     -2.007e-05  2.463e-07  -81.487  < 2e-16 ***
InterestRate                  2.999e+01  9.782e-01   30.665  < 2e-16 ***
MonthlyLoanPayment            6.828e-04  7.137e-05    9.567  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.606 on 15978 degrees of freedom
Multiple R-squared:  0.7868,    Adjusted R-squared:  0.7865
F-statistic:  2808 on 21 and 15978 DF,  p-value: < 2.2e-16
```

$$\text{BIC} = 86653.1 \qquad \text{AIC} = 86476.45$$

3. STEPWISE SELECTION

```
Coefficients:
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                    5.131e+01  6.216e-01   82.537  < 2e-16 ***
MonthlyIncome                 -1.083e-03  8.843e-06 -122.450  < 2e-16 ***
BankruptcyHistory1             1.323e+01  1.274e-01  103.856  < 2e-16 ***
DebtToIncomeRatio              1.534e+01  1.776e-01   86.359  < 2e-16 ***
NetWorth                      -2.007e-05  2.463e-07  -81.487  < 2e-16 ***
PreviousLoanDefaults1          6.804e+00  9.576e-02   71.059  < 2e-16 ***
InterestRate                   2.999e+01  9.782e-01   30.665  < 2e-16 ***
LengthOfCreditHistory         -1.697e-01  3.408e-03  -49.804  < 2e-16 ***
EmploymentStatusSelf-Employed  2.711e+00  1.071e-01   25.305  < 2e-16 ***
EmploymentStatusUnemployed     3.606e+00  1.130e-01   31.918  < 2e-16 ***
CreditCardUtilizationRate      4.895e+00  1.786e-01   27.408  < 2e-16 ***
MonthlyLoanPayment             6.828e-04  7.137e-05    9.567  < 2e-16 ***
EducationLevelBachelor        -4.511e-01  8.221e-02   -5.488 4.13e-08 ***
EducationLevelDoctorate       -1.594e+00  1.463e-01  -10.895  < 2e-16 ***
EducationLevelHigh School      3.662e-01  8.265e-02    4.431 9.44e-06 ***
EducationLevelMaster          -1.178e+00  9.784e-02  -12.040  < 2e-16 ***
Age                           -3.026e-02  2.633e-03  -11.493  < 2e-16 ***
CreditScore                   -9.754e-03  7.868e-04  -12.398  < 2e-16 ***
MonthlyDebtPayments            1.168e-03  1.187e-04    9.837  < 2e-16 ***
PaymentHistory                -3.355e-02  5.787e-03   -5.796 6.90e-09 ***
LoanAmount                     2.180e-05  3.920e-06    5.561 2.72e-08 ***
TotalLiabilities               2.659e-06  6.086e-07    4.369 1.26e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.606 on 15978 degrees of freedom
Multiple R-squared:  0.7868,    Adjusted R-squared:  0.7865
F-statistic:  2808 on 21 and 15978 DF,  p-value: < 2.2e-16
```

$$\text{BIC} = 86653.1 \qquad \text{AIC} = 86476.45$$

When applying forward, backward and stepwise selection on the training set, all three procedures converged to exactly the same model. This indicates that the resulting specification is very stable and does not depend on the particular direction of the stepwise search.

Because all selection methods chose the same set of predictors, we adopt this common specification as the final linear model for RiskScore. On its basis we generate predictions for the test set and assess the out-of-sample performance using standard error measures.

- **Mean Squared Error** = 13.14847
- **Root Mean Squared Error** = 3.626082
- **Mean Absolute Percentage Error** = 5.877203 (when calculating this error, observations with a RiskScore equal to 0 were replaced by 1 in the denominator in order to avoid division by zero).

These values indicate that the predicted RiskScore differs from the observed value on average by around 3.6 points, which corresponds to an error of about 6% of the true score. Such errors can be regarded as relatively small, so the model provides a

good approximation of the risk level on the analysed sample.

Even though the overall fit of the model is satisfactory, the diagnostic plots and Cook's distance analysis reveal the presence of a few highly influential observations. These points may distort the estimated coefficients and affect the variable selection results. Therefore, in the next subsection we investigate the impact of removing these outliers and repeat the modelling procedure on a cleaned version of the dataset in order to obtain a more robust specification for the RiskScore.

10.2 Variable selection using BIC

In this subsection, we apply the BIC criterion together with forward, backward and stepwise selection to the cleaned training dataset (after removing the most influential observations) in order to choose a suitable set of predictors for the RiskScore.

1. **FORWARD SELECTION**

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  5.131e+01  6.155e-01  83.365  < 2e-16 ***
MonthlyIncome               -8.714e-04  6.338e-05 -13.749  < 2e-16 ***
BankruptcyHistory1           1.328e+01  1.283e-01 103.511  < 2e-16 ***
DebtToIncomeRatio            1.546e+01  1.767e-01  87.510  < 2e-16 ***
NetWorth                    -2.106e-05  2.512e-07 -83.837  < 2e-16 ***
PreviousLoanDefaults1        6.732e+00  9.362e-02  71.911  < 2e-16 ***
InterestRate                 2.921e+01  9.684e-01  30.161  < 2e-16 ***
LengthOfCreditHistory       -1.698e-01  3.367e-03 -50.423  < 2e-16 ***
EmploymentStatusSelf-Employed  2.847e+00  1.067e-01  26.684  < 2e-16 ***
EmploymentStatusUnemployed   3.683e+00  1.118e-01  32.942  < 2e-16 ***
CreditCardUtilizationRate    4.683e+00  1.765e-01  26.533  < 2e-16 ***
MonthlyLoanPayment           6.340e-04  7.144e-05   8.874  < 2e-16 ***
EducationLevelBachelor      -4.967e-01  8.144e-02  -6.099 1.09e-09 ***
EducationLevelDoctorate     -1.846e+00  1.460e-01 -12.647  < 2e-16 ***
EducationLevelHigh School    3.720e-01  8.169e-02   4.554 5.30e-06 ***
EducationLevelMaster        -1.143e+00  9.688e-02 -11.803  < 2e-16 ***
Age                         -2.905e-02  2.609e-03 -11.135  < 2e-16 ***
CreditScore                 -9.482e-03  7.791e-04 -12.170  < 2e-16 ***
MonthlyDebtPayments          1.115e-03  1.172e-04   9.515  < 2e-16 ***
LoanAmount                   2.073e-05  3.859e-06   5.372 7.90e-08 ***
PaymentHistory              -2.628e-02  5.717e-03  -4.597 4.33e-06 ***
AnnualIncome                -1.780e-05  5.193e-06  -3.428 0.000610 ***
HomeOwnershipStatusOther     3.269e-01  9.807e-02   3.333 0.000861 ***
HomeOwnershipStatusOwn       4.518e-02  7.790e-02   0.580 0.561951
HomeOwnershipStatusRent      3.221e-01  6.809e-02   4.730 2.27e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.566 on 15969 degrees of freedom
Multiple R-squared:  0.7879,    Adjusted R-squared:  0.7876
F-statistic:  2472 on 24 and 15969 DF,  p-value: < 2.2e-16
```

$$\text{BIC} = 86287.68 \qquad \text{AIC} = 86088$$

2. BACKWARD SELECTION

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  5.131e+01  6.155e-01  83.365  < 2e-16 ***
Age                         -2.905e-02  2.609e-03 -11.135  < 2e-16 ***
AnnualIncome                -1.780e-05  5.193e-06  -3.428 0.000610 ***
CreditScore                 -9.482e-03  7.791e-04 -12.170  < 2e-16 ***
EmploymentStatusSelf-Employed  2.847e+00  1.067e-01  26.684  < 2e-16 ***
EmploymentStatusUnemployed   3.683e+00  1.118e-01  32.942  < 2e-16 ***
EducationLevelBachelor      -4.967e-01  8.144e-02  -6.099 1.09e-09 ***
EducationLevelDoctorate     -1.846e+00  1.460e-01 -12.647  < 2e-16 ***
EducationLevelHigh School    3.720e-01  8.169e-02   4.554 5.30e-06 ***
EducationLevelMaster        -1.143e+00  9.688e-02 -11.803  < 2e-16 ***
LoanAmount                   2.073e-05  3.859e-06   5.372 7.90e-08 ***
HomeOwnershipStatusOther     3.269e-01  9.807e-02   3.333 0.000861 ***
HomeOwnershipStatusOwn       4.518e-02  7.790e-02   0.580 0.561951
HomeOwnershipStatusRent      3.221e-01  6.809e-02   4.730 2.27e-06 ***
MonthlyDebtPayments          1.115e-03  1.172e-04   9.515  < 2e-16 ***
CreditCardUtilizationRate    4.683e+00  1.765e-01  26.533  < 2e-16 ***
DebtToIncomeRatio            1.546e+01  1.767e-01  87.510  < 2e-16 ***
BankruptcyHistory1           1.328e+01  1.283e-01 103.511  < 2e-16 ***
PreviousLoanDefaults1        6.732e+00  9.362e-02  71.911  < 2e-16 ***
PaymentHistory              -2.628e-02  5.717e-03  -4.597 4.33e-06 ***
LengthOfCreditHistory       -1.698e-01  3.367e-03 -50.423  < 2e-16 ***
MonthlyIncome               -8.714e-04  6.338e-05 -13.749  < 2e-16 ***
NetWorth                    -2.106e-05  2.512e-07 -83.837  < 2e-16 ***
InterestRate                 2.921e+01  9.684e-01  30.161  < 2e-16 ***
MonthlyLoanPayment           6.340e-04  7.144e-05   8.874  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.566 on 15969 degrees of freedom
Multiple R-squared:  0.7879,    Adjusted R-squared:  0.7876
F-statistic:  2472 on 24 and 15969 DF,  p-value: < 2.2e-16
```

$$\text{BIC} = 86287.68 \qquad \text{AIC} = 86088$$

3. STEPWISE SELECTION

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    5.131e+01  6.155e-01  83.365  < 2e-16 ***
MonthlyIncome                 -8.714e-04  6.338e-05 -13.749  < 2e-16 ***
BankruptcyHistory1             1.328e+01  1.283e-01 103.511  < 2e-16 ***
DebtToIncomeRatio              1.546e+01  1.767e-01  87.510  < 2e-16 ***
NetWorth                      -2.106e-05  2.512e-07 -83.837  < 2e-16 ***
PreviousLoanDefaults1          6.732e+00  9.362e-02  71.911  < 2e-16 ***
InterestRate                   2.921e+01  9.684e-01  30.161  < 2e-16 ***
LengthOfCreditHistory         -1.698e-01  3.367e-03 -50.423  < 2e-16 ***
EmploymentStatusSelf-Employed  2.847e+00  1.067e-01  26.684  < 2e-16 ***
EmploymentStatusUnemployed     3.683e+00  1.118e-01  32.942  < 2e-16 ***
CreditCardUtilizationRate      4.683e+00  1.765e-01  26.533  < 2e-16 ***
MonthlyLoanPayment             6.340e-04  7.144e-05   8.874  < 2e-16 ***
EducationLevelBachelor        -4.967e-01  8.144e-02  -6.099 1.09e-09 ***
EducationLevelDoctorate       -1.846e+00  1.460e-01 -12.647  < 2e-16 ***
EducationLevelHigh School      3.720e-01  8.169e-02   4.554 5.30e-06 ***
EducationLevelMaster          -1.143e+00  9.688e-02 -11.803  < 2e-16 ***
Age                           -2.905e-02  2.609e-03 -11.135  < 2e-16 ***
CreditScore                   -9.482e-03  7.791e-04 -12.170  < 2e-16 ***
MonthlyDebtPayments            1.115e-03  1.172e-04   9.515  < 2e-16 ***
LoanAmount                     2.073e-05  3.859e-06   5.372 7.90e-08 ***
PaymentHistory                -2.628e-02  5.717e-03  -4.597 4.33e-06 ***
AnnualIncome                  -1.780e-05  5.193e-06  -3.428 0.000610 ***
HomeOwnershipStatusOther       3.269e-01  9.807e-02   3.333 0.000861 ***
HomeOwnershipStatusOwn         4.518e-02  7.790e-02   0.580 0.561951
HomeOwnershipStatusRent        3.221e-01  6.809e-02   4.730 2.27e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.566 on 15969 degrees of freedom
Multiple R-squared:  0.7879,    Adjusted R-squared:  0.7876
F-statistic:  2472 on 24 and 15969 DF,  p-value: < 2.2e-16
```

$$\text{BIC} = 86287.68 \qquad \text{AIC} = 86088$$

Forward, backward and stepwise selection, applied on the training set with BIC criteria, all led to the same final model. This suggests that the chosen specification is stable and robust with respect to the selection method.

Since forward, backward and stepwise selection with the BIC criterion all produced the same set of predictors, we treat this common specification as our final linear model for RiskScore. Using this model, we produced predictions on the test set and evaluated the out-of-sample performance. As accuracy measures we computed:

- **Mean Squared Error** = 12.73798
- **Root Mean Squared Error** = 3.569031
- **Mean Absolute Percentage Error** = 5.964182 (when calculating this error, observations with a RiskScore equal to 0 were replaced by 1 in the denominator in order to avoid division by zero).

From the percentage error we see that the predictions are quite accurate. A MAPE of about 6% indicates that the selected variables provide a good approximation of the RiskScore on the test set, so this specification can be considered a reasonable choice for modelling credit risk.

10.3 Variable selection using AIC

In this subsection, we apply the AIC criterion together with forward, backward and stepwise selection to the cleaned training dataset (after removing the most influential observations) in order to choose a suitable set of predictors for the RiskScore.

1. **FORWARD SELECTION**

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  5.074e+01  6.723e-01  75.478  < 2e-16 ***
MonthlyIncome               -8.828e-04  6.377e-05 -13.844  < 2e-16 ***
BankruptcyHistory1           1.328e+01  1.282e-01 103.577  < 2e-16 ***
DebtToIncomeRatio            1.546e+01  1.766e-01  87.559  < 2e-16 ***
NetWorth                    -1.756e-05  1.528e-06 -11.495  < 2e-16 ***
PreviousLoanDefaults1        6.738e+00  9.358e-02  71.998  < 2e-16 ***
InterestRate                 3.123e+01  1.225e+00  25.488  < 2e-16 ***
LengthOfCreditHistory       -1.698e-01  3.365e-03 -50.447  < 2e-16 ***
EmploymentStatusSelf-Employed 2.844e+00  1.066e-01  26.672  < 2e-16 ***
EmploymentStatusUnemployed   3.682e+00  1.117e-01  32.951  < 2e-16 ***
CreditCardUtilizationRate    4.685e+00  1.764e-01  26.557  < 2e-16 ***
EducationLevelBachelor      -5.047e-01  8.142e-02  -6.199 5.82e-10 ***
EducationLevelDoctorate     -1.860e+00  1.459e-01 -12.752  < 2e-16 ***
EducationLevelHigh School    3.681e-01  8.166e-02   4.508 6.59e-06 ***
EducationLevelMaster        -1.156e+00  9.685e-02 -11.935  < 2e-16 ***
MonthlyLoanPayment           5.594e-04  9.365e-05   5.973 2.37e-09 ***
Age                         -2.926e-02  2.610e-03 -11.209  < 2e-16 ***
CreditScore                 -8.582e-03  8.474e-04 -10.129  < 2e-16 ***
MonthlyDebtPayments          1.177e-03  1.244e-04   9.464  < 2e-16 ***
LoanAmount                   2.394e-05  4.019e-06   5.956 2.64e-09 ***
HomeOwnershipStatusOther     3.295e-01  9.800e-02   3.363 0.000774 ***
HomeOwnershipStatusOwn       4.382e-02  7.785e-02   0.563 0.573556
HomeOwnershipStatusRent      3.223e-01  6.807e-02   4.734 2.22e-06 ***
PaymentHistory              -2.637e-02  5.714e-03  -4.615 3.97e-06 ***
AnnualIncome                -1.775e-05  5.190e-06  -3.420 0.000629 ***
TotalLiabilities             2.870e-06  7.813e-07   3.674 0.000240 ***
LoanDuration                -4.989e-03  1.834e-03  -2.720 0.006545 **
TotalAssets                 -3.330e-06  1.473e-06  -2.261 0.023791 *
SavingsAccountBalance       -6.523e-06  4.216e-06  -1.547 0.121830
NumberOfCreditInquiries      4.228e-02  2.836e-02   1.491 0.136052
JobTenure                   -1.832e-02  1.255e-02  -1.459 0.144495
TotalDebtToIncomeRatio      -2.023e-01  1.408e-01  -1.436 0.150993
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.564 on 15962 degrees of freedom
Multiple R-squared:  0.7883,    Adjusted R-squared:  0.7879
F-statistic:  1917 on 31 and 15962 DF,  p-value: < 2.2e-16
```

$$\text{BIC} = 86325.78 \qquad \text{AIC} = 86072.35$$

2. BACKWARD SELECTION

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  5.074e+01  6.723e-01  75.478  < 2e-16 ***
Age                         -2.926e-02  2.610e-03 -11.209  < 2e-16 ***
AnnualIncome                -1.775e-05  5.190e-06  -3.420 0.000629 ***
CreditScore                 -8.582e-03  8.474e-04 -10.129  < 2e-16 ***
EmploymentStatusSelf-Employed  2.844e+00  1.066e-01  26.672  < 2e-16 ***
EmploymentStatusUnemployed   3.682e+00  1.117e-01  32.951  < 2e-16 ***
EducationLevelBachelor      -5.047e-01  8.142e-02  -6.199 5.82e-10 ***
EducationLevelDoctorate     -1.860e+00  1.459e-01 -12.752  < 2e-16 ***
EducationLevelHigh School    3.681e-01  8.166e-02   4.508 6.59e-06 ***
EducationLevelMaster        -1.156e+00  9.685e-02 -11.935  < 2e-16 ***
LoanAmount                   2.394e-05  4.019e-06   5.956 2.64e-09 ***
LoanDuration                -4.989e-03  1.834e-03  -2.720 0.006545 **
HomeOwnershipStatusOther     3.295e-01  9.800e-02   3.363 0.000774 ***
HomeOwnershipStatusOwn       4.382e-02  7.785e-02   0.563 0.573556
HomeOwnershipStatusRent      3.223e-01  6.807e-02   4.734 2.22e-06 ***
MonthlyDebtPayments          1.177e-03  1.244e-04   9.464  < 2e-16 ***
CreditCardUtilizationRate    4.685e+00  1.764e-01  26.557  < 2e-16 ***
NumberOfCreditInquiries      4.228e-02  2.836e-02   1.491 0.136052
DebtToIncomeRatio            1.546e+01  1.766e-01  87.559  < 2e-16 ***
BankruptcyHistory1           1.328e+01  1.282e-01 103.577  < 2e-16 ***
PreviousLoanDefaults1        6.738e+00  9.358e-02  71.998  < 2e-16 ***
PaymentHistory              -2.637e-02  5.714e-03  -4.615 3.97e-06 ***
LengthOfCreditHistory       -1.698e-01  3.365e-03 -50.447  < 2e-16 ***
SavingsAccountBalance       -6.523e-06  4.216e-06  -1.547 0.121830
TotalAssets                 -3.330e-06  1.473e-06  -2.261 0.023791 *
TotalLiabilities             2.870e-06  7.813e-07   3.674 0.000240 ***
MonthlyIncome               -8.828e-04  6.377e-05 -13.844  < 2e-16 ***
JobTenure                   -1.832e-02  1.255e-02  -1.459 0.144495
NetWorth                    -1.756e-05  1.528e-06 -11.495  < 2e-16 ***
InterestRate                 3.123e+01  1.225e+00  25.488  < 2e-16 ***
MonthlyLoanPayment           5.594e-04  9.365e-05   5.973 2.37e-09 ***
TotalDebtToIncomeRatio      -2.023e-01  1.408e-01  -1.436 0.150993
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.564 on 15962 degrees of freedom
Multiple R-squared:  0.7883,    Adjusted R-squared:  0.7879
F-statistic:  1917 on 31 and 15962 DF,  p-value: < 2.2e-16
```

$$\text{BIC} = 86325.78 \qquad \text{AIC} = 86072.35$$

3. STEPWISE SELECTION

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    5.074e+01  6.723e-01  75.478  < 2e-16 ***
MonthlyIncome                 -8.828e-04  6.377e-05 -13.844  < 2e-16 ***
BankruptcyHistory1             1.328e+01  1.282e-01 103.577  < 2e-16 ***
DebtToIncomeRatio              1.546e+01  1.766e-01  87.559  < 2e-16 ***
NetWorth                      -1.756e-05  1.528e-06 -11.495  < 2e-16 ***
PreviousLoanDefaults1          6.738e+00  9.358e-02  71.998  < 2e-16 ***
InterestRate                   3.123e+01  1.225e+00  25.488  < 2e-16 ***
LengthOfCreditHistory         -1.698e-01  3.365e-03 -50.447  < 2e-16 ***
EmploymentStatusSelf-Employed  2.844e+00  1.066e-01  26.672  < 2e-16 ***
EmploymentStatusUnemployed     3.682e+00  1.117e-01  32.951  < 2e-16 ***
CreditCardUtilizationRate      4.685e+00  1.764e-01  26.557  < 2e-16 ***
EducationLevelBachelor        -5.047e-01  8.142e-02  -6.199 5.82e-10 ***
EducationLevelDoctorate       -1.860e+00  1.459e-01 -12.752  < 2e-16 ***
EducationLevelHigh School      3.681e-01  8.166e-02   4.508 6.59e-06 ***
EducationLevelMaster          -1.156e+00  9.685e-02 -11.935  < 2e-16 ***
MonthlyLoanPayment             5.594e-04  9.365e-05   5.973 2.37e-09 ***
Age                           -2.926e-02  2.610e-03 -11.209  < 2e-16 ***
CreditScore                   -8.582e-03  8.474e-04 -10.129  < 2e-16 ***
MonthlyDebtPayments            1.177e-03  1.244e-04   9.464  < 2e-16 ***
LoanAmount                     2.394e-05  4.019e-06   5.956 2.64e-09 ***
HomeOwnershipStatusOther       3.295e-01  9.800e-02   3.363 0.000774 ***
HomeOwnershipStatusOwn         4.382e-02  7.785e-02   0.563 0.573556
HomeOwnershipStatusRent        3.223e-01  6.807e-02   4.734 2.22e-06 ***
PaymentHistory                -2.637e-02  5.714e-03  -4.615 3.97e-06 ***
AnnualIncome                  -1.775e-05  5.190e-06  -3.420 0.000629 ***
TotalLiabilities               2.870e-06  7.813e-07   3.674 0.000240 ***
LoanDuration                  -4.989e-03  1.834e-03  -2.720 0.006545 **
TotalAssets                   -3.330e-06  1.473e-06  -2.261 0.023791 *
SavingsAccountBalance         -6.523e-06  4.216e-06  -1.547 0.121830
NumberOfCreditInquiries        4.228e-02  2.836e-02   1.491 0.136052
JobTenure                     -1.832e-02  1.255e-02  -1.459 0.144495
TotalDebtToIncomeRatio        -2.023e-01  1.408e-01  -1.436 0.150993
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.564 on 15962 degrees of freedom
Multiple R-squared:  0.7883,    Adjusted R-squared:  0.7879
F-statistic:  1917 on 31 and 15962 DF,  p-value: < 2.2e-16
```

$$\text{BIC} = 86325.78 \qquad \text{AIC} = 86072.35$$

Forward, backward and stepwise selection, applied on the training set with AIC criteria, all led to the same final model. This suggests that the chosen specification is stable and robust with respect to the selection method.

Since forward, backward and stepwise selection with the AIC criterion all produced the same set of predictors, we treat this common specification as our final linear model for RiskScore. Using this model, we produced predictions on the test set and evaluated the out-of-sample performance. As accuracy measures we computed:

- **Mean Squared Error** = 12.71045
- **Root Mean Squared Error** = 3.565172
- **Mean Absolute Percentage Error** = 5.954127 (when calculating this error, observations with a RiskScore equal to 0 were replaced by 1 in the denominator in order to avoid division by zero).

For the model obtained with AIC-based selection, the prediction errors on the test set are very similar to those of the BIC model. These values again indicate that the predicted RiskScore is on average only about 3.6 points (6%) away from the true value. The improvement over the BIC-based model is marginal, so the slightly better fit comes at the cost of a more complex specification.

10.4 Evaluation of the final RiskScore model

Comparing the models estimated on the original and on the cleaned dataset, we observe that the version without the seven influential observations achieves slightly lower MSE and RMSE, as well as smaller AIC and BIC values. This indicates a better overall fit and a more parsimonious specification. The MAPE for the cleaned model is only marginally higher than for the original one, so the loss in percentage accuracy is very small.

On balance, removing the outliers leads to a model that is more stable and easier to interpret, while maintaining a very similar (and still high) level of predictive accuracy. Therefore, in the following we base our analysis and interpretation on the RiskScore model estimated on the cleaned dataset.

Both the BIC-based and the AIC-based specifications achieve very similar prediction errors on the test set, with a low RMSE and a MAPE of around 6%. This means that, in terms of out-of-sample accuracy, the two models perform almost equally well. The AIC model has a slightly better fit, but this improvement is only marginal and comes at the cost of including more explanatory variables.

Since the BIC model is more parsimonious and easier to interpret, while providing nearly the same predictive performance as the AIC model, we treat the BIC-based specification as our final model for the RiskScore.

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    5.131e+01  6.155e-01  83.365  < 2e-16 ***
MonthlyIncome                 -8.714e-04  6.338e-05 -13.749  < 2e-16 ***
BankruptcyHistory1             1.328e+01  1.283e-01 103.511  < 2e-16 ***
DebtToIncomeRatio              1.546e+01  1.767e-01  87.510  < 2e-16 ***
NetWorth                      -2.106e-05  2.512e-07 -83.837  < 2e-16 ***
PreviousLoanDefaults1          6.732e+00  9.362e-02  71.911  < 2e-16 ***
InterestRate                   2.921e+01  9.684e-01  30.161  < 2e-16 ***
LengthOfCreditHistory         -1.698e-01  3.367e-03 -50.423  < 2e-16 ***
EmploymentStatusSelf-Employed  2.847e+00  1.067e-01  26.684  < 2e-16 ***
EmploymentStatusUnemployed     3.683e+00  1.118e-01  32.942  < 2e-16 ***
CreditCardUtilizationRate      4.683e+00  1.765e-01  26.533  < 2e-16 ***
MonthlyLoanPayment             6.340e-04  7.144e-05   8.874  < 2e-16 ***
EducationLevelBachelor        -4.967e-01  8.144e-02  -6.099 1.09e-09 ***
EducationLevelDoctorate       -1.846e+00  1.460e-01 -12.647  < 2e-16 ***
EducationLevelHigh School      3.720e-01  8.169e-02   4.554 5.30e-06 ***
EducationLevelMaster          -1.143e+00  9.688e-02 -11.803  < 2e-16 ***
Age                           -2.905e-02  2.609e-03 -11.135  < 2e-16 ***
CreditScore                   -9.482e-03  7.791e-04 -12.170  < 2e-16 ***
MonthlyDebtPayments            1.115e-03  1.172e-04   9.515  < 2e-16 ***
LoanAmount                     2.073e-05  3.859e-06   5.372 7.90e-08 ***
PaymentHistory                -2.628e-02  5.717e-03  -4.597 4.33e-06 ***
AnnualIncome                  -1.780e-05  5.193e-06  -3.428 0.000610 ***
HomeOwnershipStatusOther       3.269e-01  9.807e-02   3.333 0.000861 ***
HomeOwnershipStatusOwn         4.518e-02  7.790e-02   0.580 0.561951
HomeOwnershipStatusRent        3.221e-01  6.809e-02   4.730 2.27e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.566 on 15969 degrees of freedom
Multiple R-squared:  0.7879,    Adjusted R-squared:  0.7876
F-statistic:  2472 on 24 and 15969 DF,  p-value: < 2.2e-16
```

The selected variables are:

- MonthlyIncome
- BankruptcyHistory
- DebtToIncomeRatio
- NetWorth
- PreviousLoanDefaults
- InterestRate
- LengthOfCreditHistory
- EmploymentStatus
- CreditCardUtilizationRate
- MonthlyLoanPayment
- EducationLevel
- Age
- CreditScore
- MonthlyDebtPayments
- LoanAmount
- PaymentHistory
- AnnualIncome
- HomeOwnershipStatus

We can see that in this model almost all variables are statistically significant, as indicated by their very small p-values. The signs of the coefficients are also consistent with economic intuition. Some of the main effects can be interpreted as follows (keeping all other variables fixed):

- MonthlyIncome and AnnualIncome have negative coefficients, which means that applicants with higher income are assigned a lower RiskScore. Richer clients are therefore classified as less risky.
- DebtToIncomeRatio, MonthlyDebtPayments and LoanAmount enter the model with positive coefficients. A higher leverage and larger loan size are associated with a higher RiskScore, so more indebted applicants are assessed as riskier.

- The dummy variables BankruptcyHistory and PreviousLoanDefaults have large positive effects: clients with a past bankruptcy or previous defaults receive a substantially higher RiskScore than otherwise similar applicants without such events in their credit history
- CreditScore and NetWorth have negative coefficients. A better credit score and higher net worth reduce the predicted RiskScore, which confirms that financially stronger applicants are viewed as less risky.
- Labour-market status also matters: being unemployed or self-employed increases the RiskScore compared to regularly employed applicants, reflecting a higher perceived income risk.
- Finally, Age has a negative coefficient, indicating that older applicants tend to have a slightly lower RiskScore, while the indicators for EducationLevel and HomeOwnershipStatus capture additional differences in risk across education groups and housing situations.

Overall, the direction and significance of these coefficients are in line with credit-risk theory and support the validity of the selected linear model for the RiskScore.